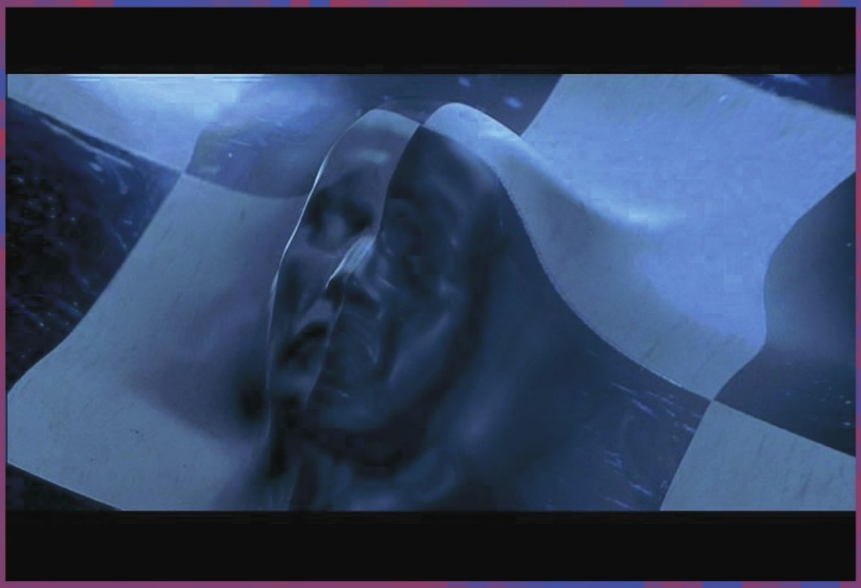
State-of-the-Art Survey

LNAI 4451

Thomas S. Huang
Anton Nijholt
Maja Pantic
Alex Pentland (Eds.)

Artifical Intelligence for Human Computing

ICMI 2006 and IJCAI 2007 International Workshops
Banff, Canada, November 2006 and
Hyderabad, India, January 2007
Revised Selected and Invited Papers



Springer

Lecture Notes in Artificial Intelligence 4451

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Thomas S. Huang Anton Nijholt
Maja Pantic Alex Pentland (Eds.)

Artifical Intelligence for Human Computing

ICMI 2006 and IJCAI 2007 International Workshops
Banff, Canada, November 3, 2006 and
Hyderabad, India, January 6, 2007
Revised Seleced and Invited Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Thomas S. Huang
Beckman Institute, University of Illinois at Urbana-Champaign
405 N. Mathews Ave., Urbana, IL 61801, USA
E-mail: huang@ifp.uiuc.edu

Anton Nijholt
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Postbus 217, 7500 AE Enschede, The Netherlands
E-mail: a.nijholt@ewi.utwente.nl

Maja Pantic
Imperial College, Computing Department
180 Queens Gate, London SW7 2AZ, U.K.
E-mail: m.pantic@imperial.ac.uk

Alex Pentland
MIT Media Laboratory, Massachusetts Insititute of Technology
Building E15, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA
E-mail: pentland@media.mit.edu

Library of Congress Control Number: 2007926703

CR Subject Classification (1998): I.2, H.5.2, I.2.10, I.4, I.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-72346-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-72346-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12059386 06/3180 5 4 3 2 1 0

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

Organization

Editorial Board

Program Committee

Editorial Assistants

Table of Contents

Foundations of Human Computing: Facial Expression and Emotion*

Jeffrey F. Cohn

Department of Psychology, University of Pittsburgh, 3137 SQ, 210 S. Bouquet Street, Pittsburgh, PA 15260 USA
jeffcohn@cs.cmu.edu
www.pitt.edu/~jeffcohn

Abstract. Many people believe that emotions and subjective feelings are one and the same and that a goal of human-centered computing is emotion recognition. The first belief is outdated; the second mistaken. For human-centered computing to succeed, a different way of thinking is needed. Emotions are species-typical patterns that evolved because of their value in addressing fundamental life tasks. Emotions consist of multiple components, of which subjective feelings may be one. They are not directly observable, but inferred from expressive behavior, self-report, physiological indicators, and context. I focus on expressive facial behavior because of its coherence with other indicators and research. Among the topics included are measurement, timing, individual differences, dyadic interaction, and inference. I propose that design and implementation of perceptual user interfaces may be better informed by considering the complexity of emotion, its various indicators, measurement, individual differences, dyadic interaction, and problems of inference.

Keywords: Emotion, measurement, facial expression, automatic facial image analysis, human-computer interaction, temporal dynamics.

1 Introduction

How can computers recognize human emotions? Is this even the correct question? By emotion, people often think of subjective feelings, but emotions are more than that and subjective feeling is in no sense essential. There is no *sin qua non* for emotion. Emotions are species-typical patterns consisting of multiple components that may include intentions, action tendencies, appraisals, other cognitions, neuromuscular and physiological changes, expressive behavior, and subjective feelings. None of these alone is necessary or sufficient for any given situation. In human-human interaction, intentions and action tendencies often are more important than what an individual may be feeling. People may or may not be aware of what they're feeling, and feelings often come about some time late in the temporal unfolding of an emotion.

* A previous version of this paper was originally published in the *Proceedings of the ACM International Conference on Multimodal Interfaces,* Banff, Canada, 2006 (Copyright © ACM Press).

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 1–16, 2007.

A goal of human-centered computing is computer systems that can unobtrusively perceive and understand human behavior in unstructured environments and respond appropriately. Much work has strived to recognize human emotions. This effort is informed by the importance of emotion to people's goals, strivings, adaptation, and quality of life [1, 2] at multiple levels of organization, from intra-personal to societal [3]. Efforts at emotion recognition, however, are inherently flawed unless one recognizes that emotion – intentions, action tendencies, appraisals and other cognitions, physiological and neuromuscular changes, and feelings – is not readily observable. Emotion can only be inferred from context, self-report, physiological indicators, and expressive behavior (see Figure 1). The focus of the current paper is on expressive behavior, in particular facial expression and approaches to its measurement, feature selection, individual differences, interpersonal regulation, and inference.

Facial expression has been a subject of keen study in behavioral science for more than a hundred years[4, 5], and within the past 10 years considerable progress has been made in automatic analysis of facial expression from digital video input [6-8].

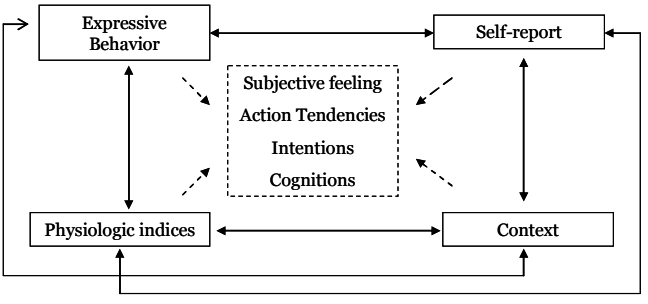


Fig. 1. Components and indicators of emotion. Solid boxes represent observables, dashed boxes latent variables. Solid arrows indicate observable correlations among indicators. Large correlations among multiple indicators indicate greater coherence among indicators. Dashed arrows represent inferential paths. Paths between emotion components are omitted. (©2006 ACM).

Facial expression correlates moderately with self-reported emotion [5] and emotion-related central and peripheral physiology [9, 10]. Facial expression has similar underlying dimensions (e.g., positive and negative affect) with self-reported emotion [11] and serves interpersonal functions by conveying communicative intent, signaling affective information in social referencing, and contributing to the regulation of social interaction [12, 13]. Expressive changes in the face are a rich source of cues about intra- and interpersonal indicators and functions of emotion [3, 14]. As a measure of trait affect and socialization, stability in facial expression emerges early in life [15]. By adulthood, stability is moderately strong, comparable to that for self-reported emotion [16].

Early work in automatic analysis and recognition of facial actions from input video focused on the relatively tractable problem of posed facial actions acquired under well-controlled conditions (e.g., frontal full-face view with minimal head motion and uniform lighting). Recent work has progressed to analysis and recognition of spontaneous facial actions with non-frontal views, small to moderate out-of-plane

head motion, subtle facial actions, and variation in illumination [17-19]. Moreover, methods of analysis and synthesis of facial expression are beginning to merge. It is becoming possible to animate an avatar from shape and appearance measures of human facial actions in real time [20], which is likely to significantly impact human-centered computing. By separating identity from facial behavior, for instance, user confidentiality could be better protected.

Here, I present key issues to consider in designing interfaces that approach the naturalness of face-to-face interaction. These include approaches to measurement, types of features, individual differences, interpersonal regulation, and inference.

2 Approaches to Measurement

Two major approaches are sign- and message judgment [21]. In message judgment, the observer's task is to make *inferences* about something underlying the facial behavior, such as emotion or personality. In measuring sign vehicles, the task is to *describe* the surface of behavior, such as when the face moves a certain way. As an example, upon seeing a smiling face, an observer with a judgment-based approach would make judgments such as "happy," whereas an observer with a sign-based approach would code the face as having an upward, oblique movement of the lip corners. Message judgment implicitly assumes that the face is an emotion "read out." Sign-based measurement is agnostic and leaves inference to higher-order decision making.

2.1 Message Judgment

Message judgment approaches define facial expressions in terms of inferred emotion. Of the various descriptors, those of Ekman have been especially influential. Ekman [22] proposed six "basic emotions." They are joy, surprise, sadness, disgust, fear, and anger. Each was hypothesized to have universally displayed and recognized signals, universal elicitors, specific patterns of physiology, rapid, unbidden onset, and brief duration, among other attributes. Of the universal signals, prototypic expressions were described for each emotion (Figure 2). Most research in automatic recognition of

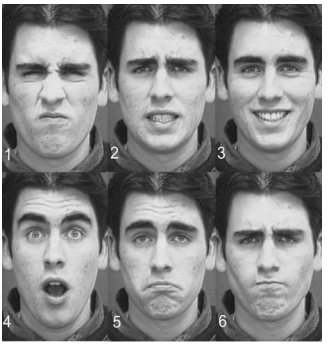


Fig. 2. Emotion-specified expressions: disgust, fear, joy, surprise, sadness, and anger. From [13]. Individual images are from the Cohn-Kanade FACS-Coded Image Database [25]. (© Jeffrey Cohn).

facial expression [26, 27] and much emotion research in psychology [28] has concentrated on one or more of these six emotions. This list, however, was never intended as exhaustive of human emotion. It is not. Rather, it was proposed in terms of conformity with specific criteria noted, such as having a universal display (i.e., prototypic expression). Other emotions that may be inferred from facial expression include embarrassment and contempt among others. Indicators of cognitive states, such as interest and confusion, have been described as well [29].

While much literature has emphasized expressions of one or another emotion, expressions may include blends or combinations of two or more [30]. For purposes such as detecting deception, expressions that include traces of contradictory emotions are of particular interest. Masking smiles[31], in which smiling is used to cover up or hide an underlying emotion are the best known. Figure 3 shows examples of two masking smiles (middle and on the right) and a "felt" or Duchenne smile (on the left) for comparison. Duchenne smiles are believed to express felt positive emotion. In the two masking smiles there is indication of sadness in one, suggested by the downward pull of the lip corners and slight pouching of the skin below them (due to AU 15 in FACS; see next section), and disgust in the other, suggested by the upward pull of the lip medial to the lip corner and philtrum and the altered shape of the nasolabial furrow (shape and appearance changes due to AU 10 in FACS) These examples come from a study in which nurses watched a movie clip intended to elicit joy and other movie clips intended to elicit disgust, revulsion, or sadness [31]. The nurses were instructed to hide their negative emotions so that an observer would be unable to tell which movie they were watching.

Fig. 3. From left to right, example of a "felt" smile, a smile "masking" disgust and one "masking" sadness. Corresponding FACS AUs are AU 6+12+26, AU 10+12+26, and AU 12+15, respectively. Reprinted with permission from [31] (© Paul Ekman).

2.2 Sign Measurement

Cohn & Ekman [32] review manual methods for labeling facial actions. Of the various methods, the Facial Action Coding System (FACS) [24, 33] is the most comprehensive, psychometrically rigorous, and widely used [32, 34]. Using FACS and viewing video-recorded facial behavior at frame rate and slow motion, coders can manually code nearly all possible facial expressions, which are decomposed into action units (AUs). Action units, with some qualifications, are the smallest visually discriminable facial movements. By comparison, other systems for measuring facial actions are less thorough and fail to differentiate between some anatomically distinct movements or consider as separable movements that are not anatomically distinct [35].

The most recent version of FACS specifies 9 action units in the upper face, 18 in the lower face, 11 for head position and movement, nine for eye position and movement, and additional descriptors for miscellaneous actions, gross body movement, and supplementary codes. (For a complete list together with anatomic basis for each AU, see [21, 36]).

Action units may occur singly or in combinations, and combinations may be additive or non-additive. In additive combinations, the appearance of each action unit is independent; whereas in non-additive combinations they modify each other's appearance. Non-additive combinations are analogous to co-articulation effects in speech, in which one phoneme modifies the sound of ones with which it is contiguous. An example of an additive combination in FACS is AU 1+2, which often occurs in surprise (along with eye widening, AU 5) and in the brow-flash greeting [37]. The combination of these two action units raises the inner (AU 1) and outer (AU 2) corners of the eyebrows and causes horizontal wrinkles to appear across the forehead. The appearance changes associated with AU 1+2 are the product of their joint actions.

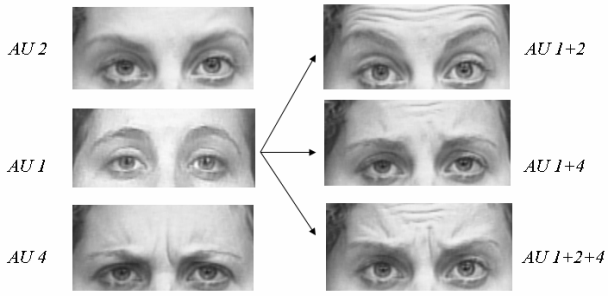


Fig. 4. Examples of individual action units and action unit combinations. AU 1+2 is an additive combination. AU 1+4 and AU 1+2+4 are non-additive, comparable to co-articulation effects in speech. (© Jeffrey Cohn).

An example of a non-additive combination is AU 1+4, which often occurs in sadness [4] (see Figure 4). When AU 1 occurs alone, the inner eyebrows are pulled upward. When AU 4 occurs alone, they are pulled together and downward (Figure 4). When AU 1 and AU 4 occur together, the downward action of AU 4 is modified. In AU 1+4, the inner eyebrows are raised and pulled together. This action typically gives an oblique shape to the brows and causes horizontal wrinkles to appear in the center of the forehead, as well as other changes in appearance that are characteristic of sadness. Automatic recognition of non-additive combinations such as this presents similar complexity to that of co-articulation effects in speech. Failure to account for non-additive combination in automatic recognition exploits the correlation among AUs and can lead to inflated estimates of algorithm performance. For further reading on FACS, see [21, 32].

2.3 Automatic Measurement

Most work in automatic analysis and recognition of facial actions has followed the message-judgment approach, with the goal of recognizing the six prototypic

expressions proposed by Ekman. Relatively few investigators have pursued the sign-based approach in which specific facial action units are recognized [7, 27]. Of the two approaches, message judgment is better suited to recognition of prototypic expressions provided the number of classes is relatively small and between-class variation relatively large. These conditions typically obtain for prototypic expressions because they are few in number and vary from each other in multiple regions of the face. Using a sign-based approach to learn AU combinations, one would have to learn first the individual muscle actions and then learn the mapping of clusters of muscle actions onto prototypic expressions. While prototypic expressions can be learned either way, the former is more efficient.

In comparison with sign-based approaches to automatic facial image analysis, the generalizability of message based approaches is more limited. Prototypic facial actions, with the exception of joy, are relatively rare in naturally occurring behavior, and when they occur, their intensity is often reduced relative to when posed. Low-intensity, or subtle, expressions are more difficult to detect for both machine learning and human observers [38]. Another issue is that emotion is more often indicated by a smaller number of facial actions than is assumed by the message judgment approach. Anger, for instance, may be communicated by slight tightening of the lower eyelids, which a message-judgment approach would be ill-prepared to detect. Masking expressions, as well, are not readily learned using message-judgment. While their component facial actions may occur with adequate frequency for learning, the specific combinations are too varied and infrequent for many machine learning approaches. With a sign-based approach, when training and testing samples for masking or other complex expressions are small, rule-based classifiers informed by human experts may be used for expression recognition. For these and other reasons, the sign-based measurement approach may prove more productive for human-centered computing.

2.4 Reliability of Meta-Data

The reliability of manually labeled images (i.e., behavioral descriptors or meta-data) is a critical concern for machine learning algorithms. If ground truth is contaminated by 20-30% error, which is not uncommon, that is a significant drag on algorithm performance. For both message judgment and sign-based approaches, similar concerns arise. Using AUs as an example, at least four types of reliability (i.e., agreement between observers) are relevant to the interpretation of substantive findings. These are reliability for occurrence/non-occurrence of individual AUs, temporal precision, intensity, and aggregates. Most research in automatic facial expression analysis has focused on occurrence/non-occurrence [7, 8].

Temporal precision refers to how closely observers agree on the timing of action units, such as when they begin or end. This level of reliability becomes important when examining features such as response latency and turn taking (see Section 5). Action unit intensity becomes important for questions such as whether facial expression is influenced by audience effects [39]. Several groups have found, for instance, that people tend to smile more intensely in social contexts than when they are alone [39, 40].

Aggregates refer to combinations of action units, which as noted may be additive or non-additive. By assessing the reliability of aggregates directly, one can more

accurately estimate both their reliability and the reliability of component action units that occur in isolation.

For each of these types of reliability, a number of metrics appear in the literature. Percentage of times that two observers agree (i.e., percent agreement) is the most-often used but least informative because it fails to correct for agreement by chance. That is, simply by knowing base rates or priors, observers are more likely to agree on behaviors that occur frequently. Chance-corrected statistics that control for base rates and guessing are more informative. Examples are Cohen's kappa (for categorical variables) and intra-class correlation (for continuous variables) [41]. In addition to average agreement or reliability across behavioral descriptors, it is informative to know the results for specific descriptors. Some behaviors are relatively easy to identify, others not. Because the reliability of manual measurement limits between-system agreement, more attention to reliability of behavioral descriptors would contribute to the success of machine learning work in automatic facial expression analysis and recognition.

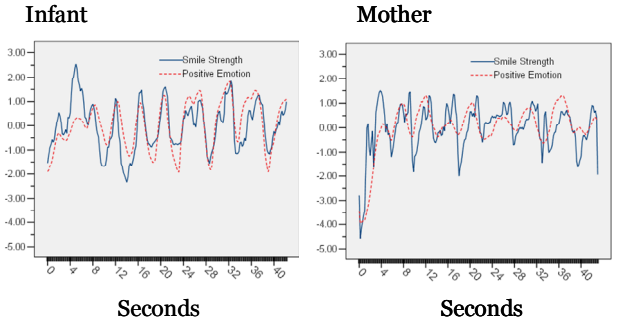


Fig. 5. Time series for AFA-measured lip-corner displacement and human-observer based ratings of positive affect in a mother-infant dyad. Data series for human observers are shifted by about ½ second to adjust for human reaction time. (©2006 ACM).

2.5 Concurrent Validity for Continuous Measurement

Most efforts at automatic expression recognition have compared inter-method agreement (i.e., concurrent validity) for categorical descriptors. Facial actions and emotion expression can vary not just in type, however, but also in amplitude or intensity. For this reason, it is important to evaluate whether alternative measurement systems have concurrent validity for continuous measures of intensity. Our research group recently examined inter-system precision for intensity contours by comparing CMU/Pitt Automatic Facial Image Analysis (AFA v.4) with continuous ratings of affective intensity and FACS intensity scoring of AU 12 by human observers. Lip-corner displacement in spontaneous smiles was measured by AFA at30 frames/second. We found high concurrent validity between the two methods (see Figure 5 for an example) [42, 43]. In other work, we found similarly high concurrent validity for intensity contours between AFA and facial EMG [40].

3 Dynamics

Both the configuration of facial features and the timing of facial actions are important in emotion expression and recognition. The configuration of facial actions (whether emotion-specified expressions or individual action units) in relation to emotion, communicative intent, and action tendencies has been the principal focus [4, 44, 45]. Less is known about the timing of facial actions, in part because manual measurement of timing is coarse and labor intensive [46]. Advances in computer-vision and graphics have made possible increased attention to the dynamics of facial expression [36, 47].

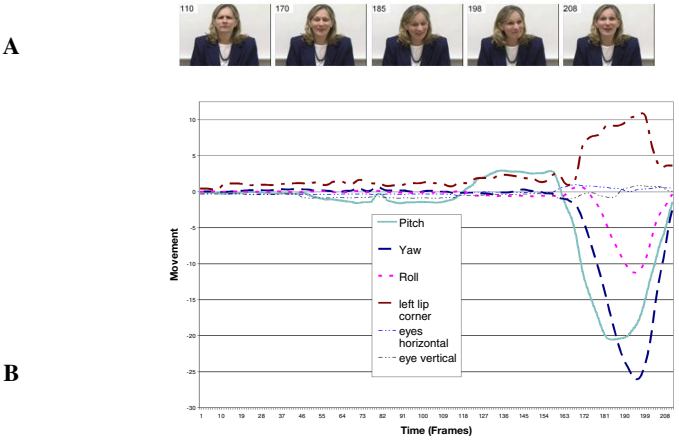


Fig. 6. Multimodal coordination of head motion, lip-corner displacement, and gaze in smiles of embarrassment. A: Selected frames from image sequence depicting embarrassment. B: Corresponding time series. Reprinted with permission from [50]. (©2004 IEEE).

There is increasing evidence that people are responsive to the timing of facial actions [48]. Consider some examples. One, facial actions having the same configuration but different timing are interpreted differently. Whether a smile is perceived as genuine or false depends in part on how quickly it changes from neutral expression to peak. Within limits, smiles with slower onset times are perceived as more genuine [49]. Two, velocity and duration of the onset phase of spontaneous smiles are highly correlated; for deliberate smiles, these components are uncorrelated. Indeed, spontaneous and deliberate smiles can be discriminated with nearly 90% accuracy from their dynamics alone [40]. Three, multimodal coordination of facial expression, direction of gaze, and head motion is a defining feature of embarrassment displays. Smile intensity increases and direction of gaze is lowered as the head pitches down and away from the partner; smile intensity decreases and gaze returns to frontal as head orientation comes back to a frontal orientation. An example is shown in Figure 6. Unless a classifier includes dynamic information, expressions cannot be accurately disambiguated.

Dynamics are critical as well to the very perception of subtle facial expressions. Subtle facial expressions that are not identifiable when viewed statically suddenly became apparent in a dynamic display [51]. An example is shown in Figure 7. Viewers were shown the same emotion expressions in one of four conditions. In the static condition, they saw only the peak expression; in multi-static, they saw each image in a video sequence from neutral to peak but with a visual mask inserted between each image. Multi-static gave viewers access to all of the images but the visual masks prevented the perception of motion. In the dynamic condition, they viewed the image sequence without the visual masks. In a first-last condition, they saw a movie sequence containing the first and last frames. Only in the latter two conditions in which motion information was afforded were expressions reliably recognized. This effect of motion was highly robust and was observed for all six basic emotions examined. An implication for machine perception is that recognition will be more difficult if performed independently for each frame, as is typically done. Dynamics are essential to human and machine perception of subtle facial expressions and to interpretation of almost all expressive actions.

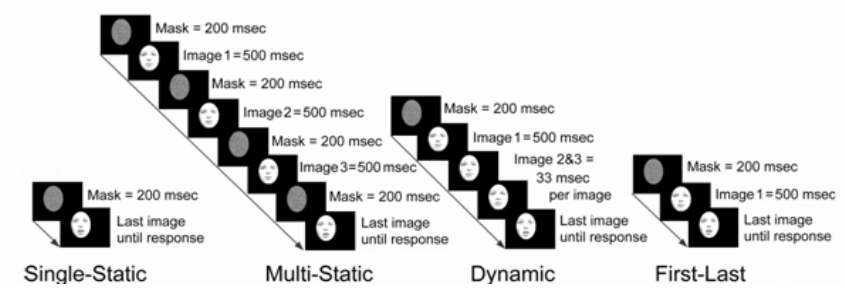


Fig. 7. Four viewing conditions: single-static, multi-static, dynamic, and first-last. Single-static shows only the final image. Multi-static shows the image sequence, but inclusion of visual masks block perception of motion. Dynamic shows the same image sequence as multi-static but includes motion information. First-last shows first and last images and motion. With some individual exceptions, subtle expressions are visible only in the latter two conditions in which motion or dynamics is visible. From [52] (©2006 APS).

When observers view facial expressions of emotion, their own faces respond rapidly and involuntarily with specific facial actions. *Zygomatic major*, which pulls the lip corners obliquely in smiling, and *corrugator supercili,* which pulls the brows together and down in frowning, contract automatically within about 0.5 seconds following perception of happy and angry faces, respectively. While such effects have been demonstrated only by facial electromyographic recordings (EMG) [53], automatic facial image analysis has recently shown strong concurrent validity with facial EMG across the range of visible movement [36]. The latter finding suggests that it is feasible using computer vision to precisely detect rapid and often low intensity increases in facial muscle activity corresponding to positive and negative emotion.

4 Individual Differences

Stable individual differences in facial expression emerge early in development and by adulthood represent 25% or more of the variation in emotion expression [16, 54]. Individual differences include reaction range for positive and negative affect, response latency, and probability of conforming to display rules. In both individual and interpersonal contexts, facial expression is moderately stable over time periods from 4- to 12 months [16] and is comparable to what has been reported previously for self-reported emotion and personality. Stability is sufficiently robust that individuals can be recognized at far above chance levels based solely on their use of facial actions (which may be considered facial behavior signatures).

In a representative study, we found moderate stability over periods of about 25 months in mothers' facial expression. Mothers were observed during face-to-face interactions with their first and second infants when each was about 2 months of age [54]. Mothers' and infants' facial expression, gaze, and vocalization were coded on a 1-s time base with behavioral descriptors. Mothers' behavior was more positive with their second-born infant (a birth-order effect), yet moderately correlated at both time points. There was no correlation between expressive behavior of first- and second-born infants; they were independent. Despite differences between infants and in mothers' interactions with each, stability in mothers' affective expression was pronounced. Findings such as these suggest that it is important to consider facial expression as bounded by range of normal variation for each person moderated by context and within which deviations may be interpreted.

Sources of individual differences in emotion expression include temperament, personality, gender, birth order (as noted above), socialization, and cultural background [55-57]. Display rules are culturally specific prescriptions for when and how to show emotion in various contexts. In some cultures, for instance, children learn not to express anger; whereas in others, anger is considered important to self expression. Among traditional Japanese, for instance, anger is less likely to be shown outside the family than in the U.S. [58]. As another example, European-American and Chinese-American couples differ in proportion of positive and negative expressions, but not in autonomic reactivity or self- reported emotion, when discussing conflicts in their relationship [59]. These differences are consistent with relative value placed on emotion expression in each culture. In Western societies, expression of positive affect is emphasized. Specific emotion expressions may also be ritualized and culture specific. The tongue-bite display (see Figure 8) communicates embarrassment/shame in parts of India and south Asia but not the U.S. Inferences about emotion become more reliable when individual differences are taken into account.

Of the various sources of individual differences, relatively little is known about the timing of facial actions. Available evidence suggests that individual differences exist here as well. Women, for instance, in addition to being more expressive than men, respond more quickly with *zygomatic major* contraction to stimuli intended to elicit positive affect [16]. When depressed, response latency is slower and more variable, which may make coordination of interpersonal timing more difficult to achieve [60]. As methods for measuring the timing of facial expression become more available, individual differences in timing will become better known.

Fig. 8. Some expressions appear in all or almost all cultures. Others are culture specific (A and B, respectively). Examples here are for embarrassment. From [61]. (©1998 Taylor & Francis).

5 Interpersonal Regulation

Interpersonal regulation includes synchrony, reciprocity, and coordinated interpersonal timing. Synchrony, or coherence, refers to the extent to which individuals are moving together in time with respect to one or more continuous output measures, such as specific facial actions, affective valence, or level of arousal. Reciprocity refers to the extent to which behavior of one individual is contingent on that of the other. Both synchrony and reciprocity have proven informative in studies of marital interaction, social development, and social psychology. Figure 9 shows an example from mother-infant interaction [42, 43]. Facial features and head motion were tracked automatically by the CMU/Pitt automated facial image analysis system version 4 [36]. The time series plot shows displacement of mother and infant lip-corners during smiles. Note that while partners tend to cycle together, there is an alternating pattern in which mother and infant take turns in leading the dyad into shared smiling.

Coordinated interpersonal timing (CIT) is the extent to which participants in a social interaction match the duration of interpersonal pauses or floor switches [62]. Floor switches are pauses that occur between the time when one person stops speaking and another begins. Coordination of floor switches follows an inverted U-shaped function in relation to affective intensity and change with development. In clinical depression, for instance, CIT becomes longer and less predictable [60] CIT has been studied most often with respect to vocal timing, but applies equally to facial expression and other modalities.

In behavioral science literature, time- and frequency domain analyses have emphasized issues of quasi-periodicity in the timing of expressive behavior and bidirectional influence with respect to amplitude (see, for instance, [63]). Lag-sequential and related hidden Markov modeling have been informative with respect to the dynamics of discrete actions and individual and dyadic states [64]. Recent work with dampened oscillator models considers regulation of changes in velocity and acceleration [65]. Most approaches assume that time series are stationary. This assumption may not always hold for behavioral data. Boker [66] identified "symmetry breaks," in which the pattern of lead-lag relationships between partners abruptly shifts. Failure to model these breaks may seriously compromise estimates of mutual influence.

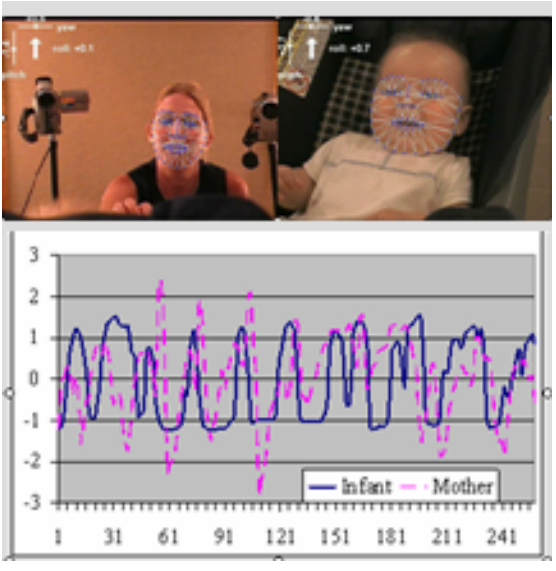


Fig. 9. Example of interaction analysis. Synchrony and reciprocity of smiling between mother and infant. Source: [42, 43]. (©2006 ACM).

6 Conclusion

Emotions are species-typical patterns that evolved because of their value in addressing fundamental life tasks [22]. They are central to human experience, yet largely beyond the comprehension of contemporary computer interfaces. Human-centered computing seeks to enable computers to unobtrusively perceive, understand, and respond appropriately to human emotion, to do so implicitly, without the need for deliberate human input. To achieve this goal, it is argued [13] that we forgo the notion of "emotion recognition" and adopt an iterative approach found in human-human interaction. In our daily interactions, we continually make inferences about other people's emotions – their intentions, action tendencies, appraisals, other cognitions, and subjective feelings – from their expressive behavior, speech, and context. The success of human-centered computing depends in part on its ability to adopt an iterative approach to inference. Computing systems are needed that can automatically detect and dynamically model a wide range of multimodal behavior from multiple persons, assess context, develop representations of individual differences, and formulate and test tentative hypotheses though the exchange of communicative signals. Part of the challenge is that the computer becomes an active agent, in turn influencing the very process it seeks to understand. Human emotions are moving targets.

Acknowledgments

Portions of this work were supported by NIMH grant R01 MH51435 and NSF HSD grant 0527397 to the University of Pittsburgh and Carnegie Mellon University.

References

1. Lazarus, R.S.: Emotion and adaptation. Oxford, NY (1991)
2. Ekman, P.: Emotions revealed. Times, New York, NY (2003)
3. Keltner, D., Haidt, J.: Social functions of emotions at multiple levels of analysis. Cognition and Emotion **13** (1999) 505-522
4. Darwin, C.: The expression of the emotions in man and animals (3rd Edition). Oxford University, New York, New York (1872/1998)
5. Ekman, P., Rosenberg, E. (eds.): What the face reveals. Oxford, New York (2005)
6. Pantic, M., Patras, I.: Dynamics of facial expressions: Recognition of facial actions and their temporal segments from profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B **36** (2006) 443-449
7. Tian, Y., Cohn, J.F., Kanade, T.: Facial expression analysis. In: Li, S.Z., Jain, A.K. (eds.): Handbook of face recognition. Springer, New York, New York (2005) 247-276
8. Pantic, M., Rothkrantz, M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1424-1445
9. Davidson, R.J., Ekman, P., Saron, C.D., Senulis, J.A., Friesen, W.V.: Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology I. Journal of Personality and Social Psychology **58** (1990) 330-341
10. Levenson, R.W., Ekman, P., Friesen, W.V.: Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology **27** (1990) 363-384
11. Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological Bulletin **98** (1985) 219-235
12. Cohn, J.F., Elmore, M.: Effects of contingent changes in mothers' affective expression on the organization of behavior in 3-month old infants. Infants Behavior and Development **11** (1988) 493-505
13. Schmidt, K.L., Cohn, J.F.: Human facial expressions as adaptations: Evolutionary perspectives in facial expression research. Yearbook of Physical Anthropology **116** (2001) 8-24
14. Gottman, J., Levenson, R., Woodin, E.: Facial expressions during marital conflict Journal of Family Communication **1** (2001) 37-57
15. Cohn, J.F., Campbell, S.B.: Influence of maternal depression on infant affect regulation. In: Cicchetti, D., Toth, S.L. (eds.): Developmental perspectives on depression. University of Rochester Press, Rochester, New York (1992) 103-130
16. Cohn, J.F., Schmidt, K.L., Gross, R., Ekman, P.: Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. International Conference on Multimodal User Interfaces, Pittsburgh, PA (2002) 491-496
17. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. IEEE International Conference on Automatic Face and Gesture Recognition, Vol. FG 2006, Southampton, England (2006) 223-228
18. Valstar, M.F., Pantic, M., Ambadar, Z., Cohn, J.F.: Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. ACM International Conference on Multimodal Interfaces, Banff, Canada (2006) 162-170
19. Lucey, S., Matthews, I., Hu, C., Ambadar, Z., De la Torre, F., Cohn, J.F.: AAM derived face representations for robust facial action recognition. Seventh IEEE International Conference on Automatic Face and Gesture Recognition, Southampton, UK (2006) 155-160

20. Boker, S.M., Cohn, J.F., Matthews, I., Ashenfelter, K., Spies, J., Brick, T., Deboeck, P., Covey, E., Tiberio, S.: Coordinated motion and facial expression in dyadic conversation. NSF Human and Social Dynamics Principal Investigators Meeting, Washington, DC. (2006)
21. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the Facial Action Coding System. In: Coan, J.A., Allen, J.B. (eds.): The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science. Oxford University, New York, NY (In press)
22. Ekman, P.: An argument for basic emotions. Cognition and Emotion **6** (1992) 169-200
23. Ekman, P., Friesen, W.V.: Unmasking the face: A guide to emotions from facial cues. Prentice-Hall Inc., Englewood Cliffs, NJ (1975)
24. Ekman, P., Friesen, W.V.: Facial action coding system. Consulting Psychologists Press, Palo Alto, CA (1978)
25. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Vol. 4, Grenoble (2000) 46-53
26. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S.: Affective multimodal human-computer interaction. ACM International Conference on Multimedia (2005) 669-676
27. Pantic, M., Rothkrantz, M.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE **91** (2003) 1371-1390
28. Keltner, D., Ekman, P.: Facial expression of emotion. In: Lewis, M., Haviland, J.M. (eds.): Handbooks of emotions. Guilford, New York (2000) 236-249
29. Scherer, K.R.: What does facial expression express? In: Strongman, K.T. (ed.): International Review of Studies on Emotion, Vol. 2. John Wiley & Sons Ltd. (1992) 138-165
30. Izard, C.E., Dougherty, L.M., Hembree, E.A.: A system for identifying affect expressions by holistic judgments. Instructional Resources Center, University of Delaware, Newark, Delaware (1983)
31. Ekman, P., Friesen, W.V., O'Sullivan, M.: Smiles when lying. Journal of Personality and Social Psychology **54** (1988) 414-420
32. Cohn, J.F., Ekman, P.: Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In: Harrigan, J.A., Rosenthal, R., Scherer, K. (eds.): Handbook of nonverbal behavior research methods in the affective sciences. Oxford, New York (2005) 9-64
33. Ekman, P., Friesen, W.V., Hager, J.C. (eds.): Facial action coding system. Research Nexus, Network Research Information, Salt Lake City, UT (2002)
34. Rosenthal, R.: Conducting judgment studies. In: Harrigan, J.A., Rosenthal, R., Scherer, K.R. (eds.): Handbook of nonverbal behavior research methods in the affective sciences. Oxford, NY (2005) 199-236
35. Oster, H., Hegley, D., Nagel, L.: Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. Developmental Psychology **28** (1992) 1115-1131
36. Cohn, J.F., Kanade, T.: Use of automated facial image analysis for measurement of emotion expression. In: Coan, J.A., Allen, J.B. (eds.): The handbook of emotion elicitation and assessment. Oxford, New York, NY (in press)
37. Eibl-Eibesfeldt, I.: Human ethology. Aldine de Gruvier, NY, NY (1989)
38. Sayette, M.A., Cohn, J.F., Wertz, J.M., Perrott, M.A., Parrott, D.J.: A psychometric evaluation of the Facial Action Coding System for assessing spontaneous expression. Journal of Nonverbal Behavior **25** (2001) 167-186

39. Fridlund, A.J., Sabini, J.P., Hedlund, L.E., Schaut, J.A., Shenker, J.J., Knauer, M.J.: Audience effects on solitary faces during imagery: Displaying to the people in your head. Journal of Nonverbal Behavior **14** (1990) 113-137
40. Cohn, J.F., Schmidt, K.L.: The timing of facial motion in posed and spontaneous smiles. International Journal of Wavelets, Multiresolution and Information Processing **2** (2004) 1-12
41. Fleiss, J.L.: Statistical methods for rates and proportions. Wiley, New York, New York (1981)
42. Messinger, D.S., Chow, S.M., Koterba, S., Hu, C., Haltigan, J.D., Wang, T., Cohn, J.F.: Continuously measured smile dynamics in infant-mother interaction. Miami (2006)
43. Ibanez, L., Messinger, D., Ambadar, Z., Cohn, J.F.: Automated measurement of infant and mother interactive smiling. American Psychological Society (2006)
44. Russell, J.A., Fernandez-Dols, J.M. (eds.): The psychology of facial expression. Cambridge University, Cambridge, United Kingdom (1997)
45. Ekman, P.: Facial expression and emotion. American Psychologist **48** (1993) 384-392
46. Cohn, J.F.: Automated analysis of the configuration and timing of facial expression. In: Ekman, P., Rosenberg, E. (eds.): What the face reveals. Oxford, New York (2005) 388-392
47. Theobald, B.J., Cohn, J.F.: Facial image synthesis. In: Sander, D., Scherer, K.R. (eds.): Oxford companion to affective sciences: An encyclopedic dictionary for the affective sciences. Oxford University Press, NY (In press) xxx-xxx
48. Edwards, K.: The face of time: Temporal cues in facial expressions of emotion. Psychological Science **9** (1998) 270-276
49. Krumhuber, E., Kappas, A.: Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. Journal of Nonverbal Behavior **29** (2005) 3-24
50. Cohn, J.F., Reed, L.I., Moriyama, T., Xiao, J., Schmidt, K.L., Ambadar, Z.: Multimodal coordination of facial action, head rotation, and eye motion. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Vol. FG'04, Seoul, Korea (2004) 645-650
51. Ambadar, Z., Schooler, J., Cohn, J.F.: Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. Psychological Science **16** (2005) 403-410
52. Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: Timing characteristics and interpretation of spontaneous smiles. (2006)
53. Dimberg, U., Thunberg, M., Grunedal, S.: Facial reactions to emotional stimuli: Automatically controlled emotional responses. Cognition and Emotion **16** (2002) 449–471
54. Moore, G.A., Cohn, J.F., Campbell, S.B.: Mothers' affective behavior with infant siblings: Stability and change. Developmental Psychology. **33** (1997) 856-860
55. Oster, H., Camras, L.A., Campos, J., Campos, R., Ujiee, T., Zhao-Lan, M., Lei, W.: The patterning of facial expressions in Chinese, Japanese, and American infants in fear- and anger- eliciting situations. Poster presented at the International Conference on Infant Studies, Providence, RI (1996)
56. Matsumoto, D., Willingham, B.: The thrill of victory and the agony of defeat: spontaneous expressions of medal winners of the 2004 Athens olympic games. Journal of Personality & Social Psychology **91** (2006) 568–581
57. Camras, L.A., Chen, Y.: Culture, ethnicity, and children's facial expressions: A study of European American, mainland Chinese, Chinese American, and adopted Chinese girls. **6** (2006) 103–114

58. Markus, H.R., Kitayama, S.: Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review **98** (1991) 224-253
59. Tsai, J.L., Levenson, R.W., McCoy, K.: Cultural and temperamental variation in emotional response. Emotion **6** (2006) 484-497
60. Zlochower, A.J., Cohn, J.F.: Vocal timing in face-to-face interaction of clinically depressed and nondepressed mothers and their 4-month-old infants. Infant Behavior and Development **19** (1996) 373-376
61. Haidt, J., Keltner, D.: Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. Emotion and Cognition **13** (1999) 225-266
62. Jaffe, J., Beebe, B., Feldstein, S., Crown, C.L., Jasnow, M.: Rhythms of dialogue in early infancy. Monographs of the Society for Research in Child Development **66** (2001)
63. Cohn, J.F., Tronick, E.Z.: Mother-Infant face-to-face interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior. Developmental Psychology **34** (1988) 386-392
64. Cohn, J.F., Tronick, E.Z.: Mother infant interaction: The sequence of dyadic states at three, six, and nine months. Developmental Psychology **23** (1987) 68 77
65. Chow, S.M., Ram, N., Boker, S.M., Fujita, F., Clore, G.C.: Emotion as a thermostat: representing emotion regulation using a damped oscillator model. Emotion **5** (2005) 208-225
66. Boker, S.M., Xu, M., Rotondo, J.L., King, K.: Windowed cross–correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods **7** (2002) 338-355

Instinctive Computing

Yang Cai

Carnegie Mellon University,
Ambient Intelligence Lab, CIC-2218
4720 Forbes Avenue, Pittsburgh, PA 15213, USA
ycai@cmu.edu

Abstract. Instinctive computing is a computational simulation of biological and cognitive instincts. It is a meta-program of life, just like universal gravity in nature. It profoundly influences how we look, feel, think, and act. If we want a computer to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even *to have* primitive instincts. In this paper, we will review the recent work in this area, the building blocks for the instinctive operating system, and potential applications. The paper proposes a 'bottom-up' approach that is focused on human basic instincts: forage, vigilance, reproduction, intuition and learning. They are the machine codes in human operating systems, where high-level programs, such as social functions can override the low-level instinct. However, instinctive computing has been always a default operation. Instinctive computing is the foundation of Ambient Intelligence as well as Empathic Computing. It is an essential part of Human Computing.

1 Introduction

What is the fundamental difference between a machine and a living creature? *Instinct!* Instincts are the internal impulses, such as hunger and sexual urges, which lead humans to fulfill these needs. Freud [1] stated that these biologically based energies are the fundamental driving forces of our life. They act everyday to protect us from danger and keep us fit and healthy. However, we are often barely aware of them.

Perhaps the most striking things for us are hidden in our cells. Recent biological studies suggest that mammalian cells indeed possess more intelligence than we can imagine [2]. For example, the cell movement is not random. It is capable of immensely complex migration patterns that are responses to unforeseeable encounters. Cells can 'see', for example, they can map the directions of near-infrared light sources in their environment and direct their movements toward them. No such 'vision' is possible without a very sophisticated signal processing system [3].

Instinctive computing is a computational simulation of biological and cognitive instincts. It actually started fifty years ago. Norbert Weiner [97] studied computational models of Gestalt, self-reproduction and learning. According to him, these functions are a part of the holistic communication between humans, animals and machine, which he called it 'Cybernetics'. In parallel, John von Neumann proposed the cellular automata model to simulate self-reproduction [4]. The model constitutes

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 17–46, 2007.

finite state cells interacting with one another in a neighborhood within a two-dimensional space. The conceptual machine is far ahead of its time. Due to the limitations in hardware, people had forgotten the idea for several decades until the 1970's: Conway rediscovered it in his article "Game of Life" [5]. In the model, an organism has its instinctual states, birth, movement, eating and death. Interesting patterns emerge from cell interactions such as blooming, oscillation or extinction. Wolfram further proves that many simple cellular interactions can produce very complex patterns, including chaos. He argues that interactive algorithms are more important than the mathematical equations [7]. The spatial and temporal interaction among entities is the key to understanding their complexity. Today, computational cellular automata have become a powerful tool to reveal the natural human algorithms, from microscopic cellular morphology [8] to mass panic movement in subway stations [9].

Instinct is a meta-program of life, just like universal gravity in nature. It profoundly influences how we look, feel, think, and act. If we want a computer to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even *to have* primitive instincts. In this paper, we will review the recent work in this area, the architecture of an instinctive operating system, and potential applications.

2 The Bottom-Up Approaches

The rapid progress of computer hardware development enables a bottom-up paradigm of computing: *embryonics* (embryonic electronics) [6]. The project BioWall is a bio-inspired hardware in size of 130 cubic feet, developed in the Swiss Federal Institute of Technology in Lausanne (EPFL). Scientists have investigated most of the possible avenues for novel computing, ranging from *phylogenetic* systems, inspired by the evolution of biological species, through *ontogenetic* systems, inspired by the development and growth of multicellular organisms, to *epigenetic* systems, inspired by the adaptation of individuals to the environment. FPGA (Field Programmable Gate Array) chips [10] were used to implement the cellular automata in the hardware. Thousands of FPGA chips were tiled together in a two-dimensional space like a wall. It is a reconfigurable parallel computer without a CPU (Center Processing Unit). This architecture allows the scientists to think outside of the box and develop novel bio-inspired software from the bottom of the thinking hardware: logic gates. Unfortunately, there is still no operating system for such a hardware implementation. Programming on this ad-hoc system could be very challenging.

Computer simulation of instinctive cognition has been studied in multiple disciplines. The computer model INTERACT mimics affective control from the viewpoint of dynamic interactionism, which combines psychological social psychological, system theory, linguistic, attitude measurement and mathematical modeling [11]. The computer model DAYDREAMER mimics the human daydreaming process and provides an empirical tool to study the subconscious programs embedded inside our mind [12].

Pentland coined the term 'Perceptual Intelligence' for the intelligence without words or logic [13]. He argues that if we can give computers the ability to perceive,

classify and anticipate human interactions in a human-like manner, then the computer will be able to work with humans in a natural and commonsense manner. To understand and predict human behavior, the team has developed vision algorithms to augment faces, emotions and gestures. The work is then extended to the study of 'Human Dynamics' for a broader range of social interactions, such as mobile communication. This inspires a new wave of 'Human Computing' [104] studies. Instinctive Computing is an essential part of 'Human Computing', which focuses on low-level functions and needs, similar to the machine code to software. While Human Computing algorithms attempt to answer questions such as, *who* the person is, *what* is communicated, *where* is the location, *when* did it happen, and *how* did the information get passed on. Instinctive Computing, on the other hand, attempts to answer the question of *why* someone behaves that way and predicts the consequences of one's behavior using commonsense.

Indeed *instinct is commonsense.* It has been an immense challenge to AI. For over 20 years, with over a 20-million-dollar investment, Lenat and his colleagues have been developing Cyc, a project that aims to create a reusable general knowledge base for intelligent assistants [14]. Cyc essentially is an ontological representation of human consensual knowledge, which can construct a semantic web where meaning is made explicit, allowing computers to process information intelligently. However, even Cyc has not touch the inner layers of human instincts.

Instinct often controls emotion. If we trace our anger, fear and, sadness back to its origin, we always can find the signatures of our instincts. The latest scientific findings indicate that emotions play an essential role in decision-making, perception, learning and action. They influence the very mechanisms of rational thinking. According to Picard [15], if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even *to have* and express emotions. Minsky articulates emotion as a resource-intensive process [16]. In addition, he outlines a six-level model of mind from top to bottom: self-conscious reflection, self-reflection thinking, reflective thinking, deliberative thinking, learned reactions, and instinctive reactions. He elaborates the instinctive reactions are a set of "if-then" rules, which are primitive model of instinctive computing. According to Minsky, how to sort out the priority of such rules is quite a challenge.

Instinct has an impact on how we look at things. Researchers have studied Aesthetic Computing for understanding forms, functions, dynamics and values. Cohen seeks the answer to what are the minimal visual elements to be an interesting form [17]. He also ponders how to mix the paint color without perception. Leyton argues that shape is the memory storage [18]. By instinct, we use the imaginary symmetric structure to remember the shape process.

Instinct constitutes our social behaviors as well. Computational studies of human dynamics have been rapidly growing in recent decades. At the macroscopic level, the fundamental studies in social networks such as "the six degrees of separation" [19] and the "power law of the linked interactions" [20] shed lights on the scalability of human networking activities. Those remarkable models enrich our in-depth understanding of the dynamics in a very large network, which is a challenge to a visualization system. Spectrum-graph [21] is used to visualize human activities from ambient data sources such as gas stations and cellular phone towers. Graph models

such as minimal graph cuts provide abstract, yet visual tools for analyzing the outliers in a very large social network [22]. Stochastic process based geographical profiling models have been developed to investigate serial killer's spatio-temporal patterns from the collected field data [23]. Furthermore, the cellular automata based panic model simulates the mass dynamics in public places such as train stations. The method computationally incorporates modeling, rules and visualization in one algorithm, which enables pattern discovery and rapid empirical experiments [24]. In a nutshell, the paradigm of the visual analytic social networks has been shifted from merely visual data rendering to model-based visual analysis. However, a single model may not be a panacea as many researchers have claimed. The adaptability and interactions between models and visual interfaces are perhaps potential solutions.

Computational visualization of human dynamics has been growing exponentially. Decades ago, human motion studies were largely dependent on the time-lapped photography, where joints were highlighted to form motion trajectories. Today, digital motion capturing and modeling systems enable the high fidelity modeling of the motion characteristics. Functional MRI (fMRI) systems visualize human nerve and cardiac dynamics in real-time, which has revolutionized the way of physiological and psychological studies such as driving [25]. Artificial Intelligence computational models also provide augmented cognition behaviors in navigation, planning and problem solving. A driver's eye gazing model [26], for example, is based on the classic ACT-R model [27]. The model mimics a human driver's visual experience in a rule-based system. However, as the developers of the system pointed out, how to incorporate the sensory interactions in the rule-based model is a challenge. Traditional AI models such as sequential and parallel processing have not been able to simulate the emergent behaviors.

3 Building Blocks of Instinctive Computing

If we compare to our instinctive functions with a computer operating system, we would find some similarities: both are low-level programs (machine code). They intimately deal with physical devices; they operate autonomously as default. Their programs can be 'burned' into a ROM (Read-Only Memory). They can be overridden by high-level programs (applications). Besides, they all have very limited resources.

In this study, we view instinct as a metaphor for a future computer operating system, which includes the meta-programs for forage, vigilance, reproduction, intuition, and learning.

Forage is a basic urge for us to sustain life. Computational Swarm Intelligence [28] simulates ants' navigation process based on the trace of their pheromone. The algorithms have been widely adopted into network optimization applications, such as Traveling Salesman Problem, ad-hoc wireless mesh networking and network flow planning [28]. At the digital age, information foraging has become critical. The good news is that we can fetch valuable information with just a few mouse clicks. The bed news is that spyware or malware try to harvest sensitive data such as email addresses, personal identification numbers, birthdays and passwords, and documents. Bad interactions are real interactions. They contribute variety to a digital ecosystem, where novel algorithms emerge, for example, 'honey pots' and 'food-chains'.

Vigilance comes from our fears. A death instinct is universal to all creatures. Alarm pheromones are released by creatures such as fish and bees when they alert others of danger [29-31]. Although human alarm pheromones are still debatable, there is no doubt that our instinct often makes us aware of dangerous situations. Our tongue has evolved to have 5,000 taste buds - letting us know what to swallow, and what to spit out [32]. And we also have an instinctive reaction to things which could give us a disease or make us sick. Our feelings of disgust have helped keep us safe for hundreds of generations. People can usually sense trouble with a car from noises, vibrations, or smells. An experienced driver can even tell where the problem is. Instinctive computing aims to detect anomalous events from seemingly disconnected ambient data that we take for granted. For example, the human body is a rich ambient data source: temperature, pulses, gestures, sound, forces, moisture, et al. Many electronic devices also provide pervasive ambient data streams, such as mobile phones, surveillance cameras, satellite images, personal data assistants, wireless networks and so on.

Intuition is a subconscious perception without thinking. Drawing portraits upside down allows novice artists to reproduce lower-level image features, e.g., contours, while reducing interference from higher-level face cognition [33]. A study shows that our subconscious intuition detects shape anomaly more accurately than conscious judgment [34]. Human-like perception systems have potential for remote sensing, virtual reality, medical discovery, autonomous space exploration and artificial organs that extend our perception. The peripheral vision of the redundant information enables us to detect anomalies from seemingly disconnected ambient data that we take for granted. Robertsson, et al [35] developed artificial sensor models for olfaction and manipulation, which enable knowledge discovery in a sensor web. Krepki and his colleagues [36] developed the Berlin Brain-Computer Interface (BBCI) as a potential HCI channel for knowledge discovery. Derived from an interface for physically challenged people, the brain-computer interface enables information retrieval directly through the human brain. In addition, it provides feedback regarding human attention, interests and emotion directly to an integrated computer. For decades, information analysts have been searching for ways to incorporate an expert's preference, style, attention, rhythm and other properties of intelligence into a knowledge discovery system. Unfortunately, most existing user-modeling methods are both invasive and indirect. A brain-computer interface shows us a potentially ambient approach to solving the problem.

Reproduction is a means for immortality. The instinct to have sex is one of the most potent we possess. It is vital if we are to produce the next generation. It has a great impact on the way we look, the way we smell and what we possess, which can attract the ideal mate [32]. Computational self-reproduction has been studied for half of a century. Von Neumann proposed a cellular automata to simulate the process [37]. However, so far, most of computational models are asexual. Today, online sexual interaction pushes technologies to the edge. Studies about sexual objects and interaction emerged, i.e., the computer vision model for detecting nude figures in a picture [38].

Learning upgrades instinctive programs. In 1896, James Mark Baldwin proposed that individual learning can explain evolutionary phenomena that appear to require inheritance of acquired characteristics [39]. The ability of individuals to learn can

guide the evolutionary process. Baldwin further proposed that abilities that initially require learning are eventually replaced by the evolution of genetically determined systems that do not require learning. Thus learned behaviors may become instinctive behaviors in subsequent generations, without appealing to inheritance.

In the following sections, we will review the case studies for virtual forage, vigilance, intuition, reproduction and learning.

4 Virtual Forage

Kurt Letwin said: "Our behavior is purposeful; we live in a psychological reality or life space that includes not only those parts of our physical and social environment to us but also imagined states that do not currently exist." Today, ambient information is collected everywhere, from our shopping habits to web browsing behaviors, from the calls between businesses to the medical records of individuals. Data is acquired, stored and gradually linked together. In this massive data there are many relationships that are not due to chance, but transforming data into information is not a trivial task. Data is obtained from observation and measurement and has by itself little intrinsic value. But from it we can create information: theories and relationships that describe the relationships between observations. From information we can create knowledge about high-level descriptions of what and why, explaining and understanding the fundamental data observations.

4.1 Information Foraging

Many great discoveries in history were made by accident and sagacity. True serendipity emerges from random encounters, such as in daydreaming [40-42]. In Beale's study [43], an intelligent system was designed to maximize pseudo-serendipity [44], which describes accidental discoveries of ways to achieve a desired goal. Beale introduces a synergistic interaction scheme that includes interactive data mining and a novel genetic algorithm to support serendipitous discoveries. Beale intends to answer questions such as: "what is interesting?" and "what is surprising?" In Beale's study, the high dimensional data are mapped to a visual space where data are clustered by pseudo-physics properties such as mass-spring relations. This method allows the user to interact with the data space from different perspectives and hypotheses.

Analytical models intend to reveal inner structure, dynamics or relationship of things. However, they are not necessary intuitive to humans. Conventional scientific visualization methods are intuitive but limited by dimensions and resolutions. To bridge the gap, transformation algorithms are designed to map the data from an abstract space to an intuitive one. For example, a spectrogram maps an invisible sound to a visible frequency-intensity-time space. The convergence of scientific visualization and data mining creates a new domain for visual data mining. Seeing and understanding together enable humans to discover knowledge and deeper insight from a large amount of data [45]. This approach integrates the human's Ambient Intelligence with analytical computation to form a coherent knowledge discovery environment.

4.2 Virtual Food-Chains

We all work along a metaphor of 'food chains'. Modern organizations hunger for resources and profit. In a factory, a manager is responsible to his boss and clients (providers). On the other hand, the manager is also a provider to his staffs and suppliers. The manager has to maximize his time to correspond with his boss and clients on a daily basis, while minimizing his time dealing with the problems from suppliers and staffs. Fig. 1 illustrates the mental map of the 'food chains' of a factory manager.

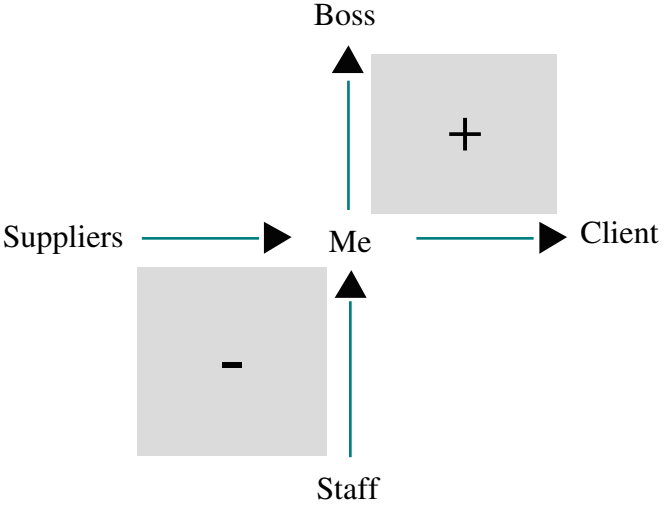


Fig. 1. Graphical user interface inspired by the food chains at workplace, where the user positions his boss on his North, client on East, staff on South and suppliers on West. The mental map helps him to organize daily business.

Unfortunately, modern electronic devices overwhelm us with information. For example, emails on the handheld PDA Blackberry™ are arranged in a linear order. Browsing through a long list of emails on a mobile phone at an airport can be miserable. Why don't we design a graphical interface that projects an individual's cognitive map? Given a sequence of messages, we can translate the message identifications into a two-dimensional quad-space, where supply-demand relations are defined as food chains, where the user positions his boss on his North, client on East, staff on South and suppliers on West. The mental map helps him to organize daily business.

4.3 Near-Field Interactions

Wireless devices enable users to interact with local information services, such as Internet, vending machines, cash machines, and check-out desks, normally within 20 meters. A wireless local network, for example, can provide a serendipitous user positioning system. Based on the Radio Signal Strength Indication (RSSI) in an

indoor wireless environment, the positioning system can estimate the distance between the access point and the wireless device [46]. The triangulation of the user's location can be calculated from multiple access points. However, in many cases, only one access point is actively connected. Indoor furniture information and the Bayesian model are used to improve positioning accuracy with physical constraints and historical ground truth data.

Fig. 2 shows a screen capture of the wireless laptop positioning output. It shows the mobile users work near the wireless access points and power supplies. It is a novel tool to study the human dynamics from ambient data. Combined with multi-modal sensors, such as infrared and magnetic signals, the positioning accuracy can be further improved. The widely distributed open sensory system also raises serious concerns about data privacy [47]. Fig. 2 shows an output from a wireless device positioning system at a building, where the location of wireless users and access points are visible on the Web. The identity of users is replaced with a dot to preserve individual privacy.

Fig. 2. Left: CMUSky wireless network device location system. The yellow dots show the dynamic patterns of mobile users. Right: Saharan ants made interesting forage patterns.

5 Vigilance of Vulnerability

Network security is vital to our economy and personal rights. Detecting the vulnerability of a network in real-time involves a massive data space and intensive processing. Can we use instinctive computing to detect the vulnerability and anomaly?

We have developed a general visualization system for rendering at least 10,000 points and each point has at least 64 attributes in real time. An important part of the pre-processing task is to normalize the continuous data, so each attribute can have the same level of influence when comparing one data to another (calculating the Euclidean distance).

The normalization is performed dividing each attribute by the maximum attribute's value of the whole data scanned during a period of time. To visualize each network connection and their 41 attributes in 2D, we use the star glyphs, where the dimensions are represented as equal-spaced angles from the center of a circle and each axis

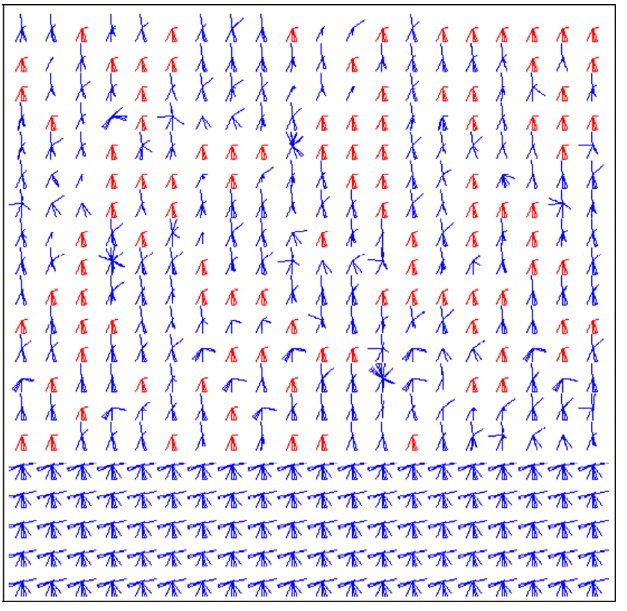


Fig. 3. This shows 400 network connections displayed on a Glyph form. The glyphs highlighted in red are connections that have similar attributes and consequently similar glyph's form. The glyphs on the bottom are connections from a DoS (Denial of Service) attack and, comparing to all the other connections, it is easy to remark that they present an abnormal form.

represents the value of the dimension. Fig. 3 shows 400 network connections displayed on a Glyph form. The glyphs highlighted on red are connections that have similar attributes and consequently similar glyph's form. The glyphs on the bottom are connections from a DoS (Denial of Service) attack and, comparing to all the other connections, it is easy to remark that they present an abnormal form.

Once the clusters are created, it is necessary to organize them, so they can be dispersed in a proper way, making an easy visualization. Several methods can be used: Principal Component Analysis, Kohonen Self Organizing Maps (SOMS) [48] and Multidimensional Scaling (MDS) [49-50]. Both SOMS and MDS were tested on the algorithm. The MDS presented a faster computational time. That is why it was chosen for the algorithm.

The goal of the MDS algorithm is to organize a multidimensional data on a two dimensional graphic by coordinate pairs (x,y). The Cartesian plane makes axes explicit and it is easier to organize data according to their characteristics. The idea of the MDS is to measure all the data distance in an n-dimensional space and place it on 2D display so they obey the same mutual distance relationship. However, a perfect 2D configuration is not always possible. Let d_{ij} be the multidimensional distance between a point i and j, calculated with the Euclidean distance. Let also δ_{ij} be the 2D distance between the same point i and j calculated with the Pythagorean Theorem, $\delta_{ij} = \sqrt{x^2 + y^2}$. If $d_{ij} \neq \delta_{ij}$, then there is a stress between the data and 2D representation. The stress is minimized with a Simplex optimization algorithm [51].

In the Simplex algorithm, first, we randomize an n-vector of independent variables as the initial point to calculate the stress. Then, it moves the point of the Simplex where the stress is largest through the opposite face of the Simplex to a lower point. Such reflections are constructed to conserve the volume of the simplex, then it will maintain its nondegeneracy. When it can do so, the method expands the Simplex in one or another direction to take larger steps. When it reaches a "valley floor," the method contracts itself in the transverse direction and tries to ooze down the valley. If there is a situation where the Simplex is trying to pass it, it contracts itself in all directions, pulling itself in around its lowest point. At this point, we get the minimum stress.

$$stress = \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} d_{ij}^2}$$

Because of its complexity, the program splits the calculation into some phases. In every phase, it will calculate the minimum stress and organize them accordingly until it has reached the globally minimum stress. Thus, it is possible to watch the glyph's movements finding the best organization among them. Once the equation above is minimized, the data positions are set and organized, respecting its distance relationship with a minimum error.

The clusters organize the data dynamically, providing to them an approximate position next to the cluster associated to them. Fig. 4 displays an example of the implemented algorithm. The figure shows the MDS organization of the clusters in form of glyphs and a zoom of a cluster (in red), where it is possible to visualize the data associated to it.

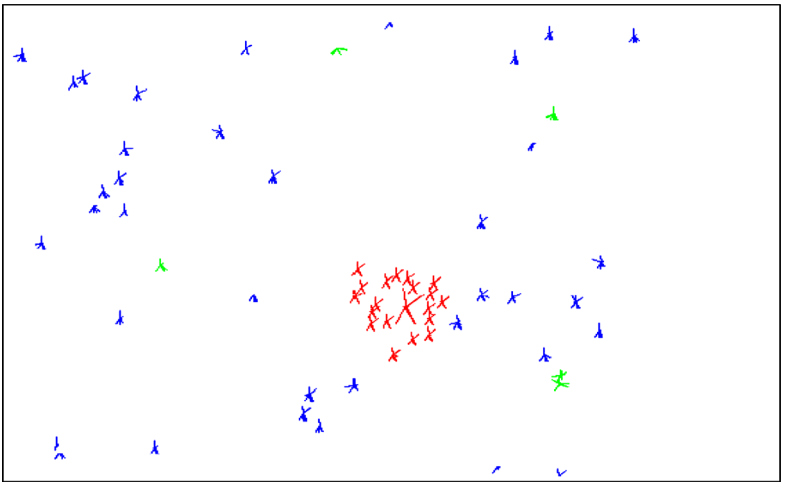


Fig. 4. Incrementally clustering. The glyphs in red are the data from a highlighted cluster. The glyphs in blue and green are the other clusters organized with the MDS algorithm.

The program supports a real-time application and the MDS organization algorithm runs in an on-the-job mode. Each cluster, here, is represented by a blue glyph. If users want to observe each data from the selected cluster, the application also provides an automatic "zoom in" every time the mouse is clicked on one of the clusters. Besides that, the tool also can be displayed in a multi-monitor mode, where different types of data can be placed in different monitors.

Human vision has about 0.1 second vision latency [98] which has been an important factor in modern video and movie technologies. In principle, a system need not update data faster than human's response time. In light of this, we can use the human latency to design many novel human-centric computing algorithms that incorporate the latency factor. Many visualization methods involve time-consuming algorithms for clustering and optimization. Instead of waiting for minutes to see the updated results, the latency-aware algorithms are designed to synchronize with the speed of human vision by incremental updating. This should create a new philosophy of algorithm design.

We tested the models with the data from KDD CUP 1999 [52]. This data represents thousands of connections with 41 different attributes, including continuous and discrete attributes. The system successfully demonstrated its capability of rendering the anomalous events, such as the DoD attacks, in both stationary and dynamic visualization.

6 Human-Like Sensors

Human sensors are great inspirations for designing an efficient perception and communication system. They can also help us to diagnose diseases from ambient information.

6.1 Multi-resolution Sensing

We have surprisingly low visual acuity in peripheral vision (rods) but very high visual acuity in the center of gaze (cones). Curiously, despite the vitality of cones to our vision, we have 125 million rods and only 6 million cones. Our gaze vision is optimized for fine details, and our peripheral vision is optimized for coarser information. Human information processing follows the power law. If the data are plotted with the axes being logarithmic, the points would be close to a single straight line. Humans process only a very small amount of information in high fidelity, but large amounts of information in middle or low fidelity. The amount of processed information is roughly inversely proportional to its level of fidelity. Therefore, we have a fidelity power law for assigning the information processing capability for sensory channels. Given the amount of sensory information X, and the processing fidelity Y, constants a and b, the relation can be expressed as:

$$Y = -a \cdot log(X) + b$$

Considering a remote sensing system for monitoring a community of elderly people, how many screens do we need for the control room? How many operators do we need for vigilance around the clock? In author's recent study [53], eye gaze tracking and face detection technologies are applied to optimize the throughput of a wireless

mobile video network. The objective of this task is to formulate a feedback control model, where the network traffic is a function of the visual attention. Given a number of camera live video channels with adjustable resolutions high and low that are arranged on a monitor screen, find the minimal network traffic as the computer detects which video channel is selected.

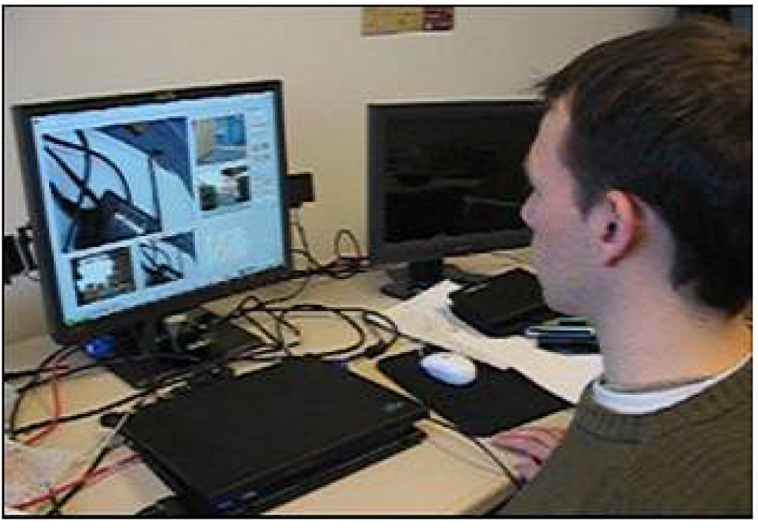


Fig. 5. The eye gaze tracking system for multiple video resolutions

To collect the lab data, the system needs to sense an operator's attention ubiquitously by tracking user's eye gaze on the monitor. Researchers at Ambient Intelligence Lab use an eye-tracking device with two infrared light sources. The camera can capture the eye gaze at 60 frames per second and the accuracy of 3 degrees. In the experiment, the operator uses their eyes to switch the video channels from a low resolution to a high resolution. The traffic monitor software records the real-time data flow.

From the empirical experiments it is found that multiple resolution screen switching can reduce the network traffic by 39%. With eye gazing interface, the throughput of the network reduced about 75%. Combining an eye tracking and face detection in the video, the overall throughput reduction reaches about 88%.

6.2 Tongue Inspection

For over two thousand years, physical inspection has been a unique and important diagnostic method of Traditional Chinese Medicine (TCM). Observing abnormal changes in the tongue, blood volume, pulse patterns, breath smells, gestures, etc., can aid in diagnosing diseases [54, 56]. TCM diagnosis is a black-box approach that involves only input and output data around the body. For many years, scientists have been trying to use modern technologies to unleash this ancient knowledge base. For example, the computer-based arterial blood-volume pulse analyzer is a 'rediscovery' of the diagnostic method originated from ancient TCM [55].

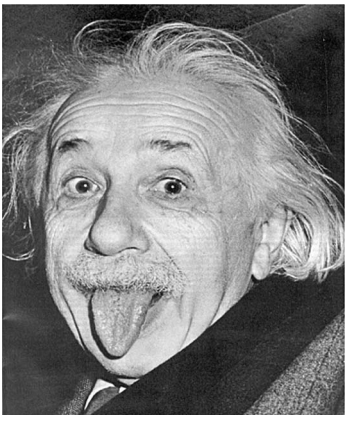


Fig. 6. According to the medical doctor, professor and author Claus Schnorrenberger from German Research Institute of Chinese Medicine, Einstein's tongue[1] reveals that he has probably suffered from insomnia. He may have been affected by a stomach disorder and constipation [56].

Visual inspection of the tongue has been a unique and important diagnostic method of Traditional Chinese Medicine (TCM) for thousands of years. The inspection of the tongue comprises the inspection of the tongue body and the coating. In the study [57], the author uses a portable digital scanner to acquire the tongue image. The features on the tongue are represented as a vector of variables such as color space coordinates L*a*b, texture energy, entropy and fractal index, as well as crack index. With a Probability Neural Network, the model reveals the correlation between the colon polyps and the features on the tongue.

The study shows the potential of inexpensive mobile cameras playing a role in healthcare. TCM diagnosis is not a replacement of the modern diagnostic technologies such as MRI, CT, Ultrasound, DNA, but an alternative tool for an early warning that brings people for further clinical diagnoses. With the growing digital technologies, it is possible to see more personal diagnostic tools in stores, just like those pregnancy test kits or diabetes self-test kits today.

7 Reproductive Aesthetics

The instinct to have sex is one of the most potent we possess. It's vital if we are to produce the next generation. It has a great impact on the way we look, the way we smell and what we possess, that can attract the ideal mate [101].

Computational self-reproduction has been studied for half of a century. John von Neumann [37] proposed cellular automata to simulate the process. However, so far, most of computational models are asexual and mechanical. Online sexual interaction pushes technology to its edge. Studies about sexual objects and interaction emerged. A computer vision model has been developed for detecting nude figures in a picture [38].

[1] Albert Einstein stuck his tongue out at obtrusive press people and paparazzi on his 72nd birthday, March 14, 1951. He once stated, "A man's interactions with his fellow man are often not so different from a dog's interactions with a fellow dog"[100].

7.1 Sexuality

The English writer William Shenstone once said: "Health is beauty, and the most perfect health is the most perfect beauty." Why do people like symmetric faces and bodies? A simple explanation: symmetric shapes indicate fitness and good health.

In every century, three body parts - breasts, waists and thighs - are more often referred to as more beautiful than other body parts. Psychologists Devendra Singh and Adrian Singh show that men have only one thing on their minds: a woman's WHR - waist-hip ratio, calculated by dividing waist circumference by that of the hips. In the Royal Society journal, Proceedings of the Royal Society, Biological Sciences [58], they analyze thousands of examples from British literature from the 16th to 18th centuries with Peter Renn of Harvard University to show that what men find attractive today was also true hundreds of years ago: a narrow waist and thus an hourglass shape.

Modern science reveals that an hourglass shape in women is associated with relatively high levels of the hormone estrogen. Since estrogen levels influence fertility, men seeking to pass their genes to the next generation would do well to pick hourglass-shaped women. As a corollary, a sizeable belly is reliably linked to decreased estrogen, reduced fecundity and increased risk for major diseases according to research conducted over the past decade.

7.2 Detecting Human Body Features

From a computer vision point of view, detecting features from 3D body scan data is nontrivial because human bodies are flexible and diversified. Function fitting has been used for extracting special landmarks, such as ankle joints, from 3D body scan data [59-60], similar to the method for extracting special points on terrain [61]. Curvature calculation is also introduced from other fields such as the sequence dependent DNA curvature [38]. These curvature calculations use methods such as chain code [62], circle fit, ratio of end to end distance to contour length, ratio of moments of inertia, and cumulative and successive bending angles. Curvature values are calculated from the data by fitting a quadratic surface over a square window and calculating directional derivatives of this surface. Sensitivity to the data noise is a major problem in both function fitting and curvature calculation methods because typical 3D scan data contains loud noises. Template matching appears to be a promising method because it is invariant to the coordinate system [59-60]. However, how to define a template and where to match the template is challenging and unique to each particular feature. In summary, there are two major obstacles in this study: robustness and speed. Many machine learning algorithms are coordinate-dependent and limited by the training data space.

An Analogia (Greek: αναλογία, means ‘proportion’) Graph is an abstraction of a proportion-preserving mapping of a shape. Assume a connected non-rigid graph G, there is an edge with a length u. The rest of the edges in G can be normalized as $p_i = v_i / u$. Let X and Y be metric spaces d_X and d_Y. A map $f: X \rightarrow Y$ is called Analogia Graph if for any x,y∈ X one has $d_Y(f(x),f(y))/u = d_X(x,y)/u$.

Use of methods similar to Analogia Graph is common in the arts. The Russian Realism painter Ropin said that the secret of painting is “comparison, comparison and

comparison." To represent objects in a picture realistically, a painter has to constantly measure and adjust the relationship among objects. "You should use the compass in your eyes, but in your hands," Ropin said. Instead of using absolute measurement of the distances and sizes, artists often use intrinsic landmarks inside the scene to estimate the relationships. For example, using numbers of heads to estimate the height of a person and using length of the eyes to measure the length of a nose, and so on. Fig. 7 is an Analogia Graph of a human body.

Using this artistic approach, we can create a graph where nodes represent regions and are connected to each other by edges, where the weight is defined as the distance between the nodes in proportion to the height of the head. Initially, we stretch the graph such that it overlays the entire body. We then create a link between each node and its respective counterpart. We link the head, shoulders, arms, elbows, hands, neck, breasts, waist, legs, knees, and feet to their respective regions. There is some tweaking required to assure that the waist region does indeed cover that area. Here we run a quick top-down search through the plane slices until there is at least two disjoint areas, which we consider to be the middle of the waist. This change also makes modifications to where the knees and breasts are, and how large their regions are.

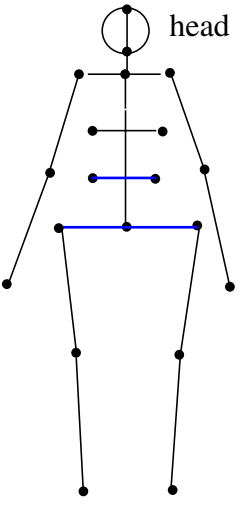


Fig. 7. Analogia graph of a human figure

We take into account that not every subject has all four limbs. Our algorithm still accepts the scan if such items are missing, such as half an arm or half a leg. It is also amenable to a complete loss of an arm or leg by looking at the expected ratio versus the real ratios when determining the length of each particular region.

However convenient it is to find such broad range of regions, it is not possible to expand this algorithm to find more details like specific fingers, toes, ankle joints, or the nose. These searches are more complicated and require additional template fitting per feature and would significantly reduce the algorithm's run time.

We found that the intrinsic proportion method can reduce the search space by an order of magnitude. In addition, it reduces the risk of finding the local optima while searching the whole body.

Intrinsic proportion measurements have been used in architecture and art for thousands of years. Roman architect Vitruvius said that the proportions of a building should correspond to those of a person, and laid down what he considered to be the relative measurements of an ideal human. Similarly in art, the proportions of the human body in a statue or painting have a direct effect on the creation of the human figure. Artists use analogous measurements that are invariant to coordinate systems. For example, using the head to measure the height and width of a human body, and using an eye to measure the height and width of a face.

Fig. 8 shows the distribution of head to body proportions calculated from the CAESAR database. The results show that on average a human is six to eight heads tall. Based on our observations from one hundred 3D scan data sets of adults from sixteen to sixty-five years old, including subjects from North America, Europe and Asia, we found that the length of one and a half head units from the bottom of the head is enough to cover the chest area. In addition, the chest width is about three heads wide. The chart on the right of Fig. 8 shows an output from the intrinsic proportion calculation based on the sample from CAESAR database.

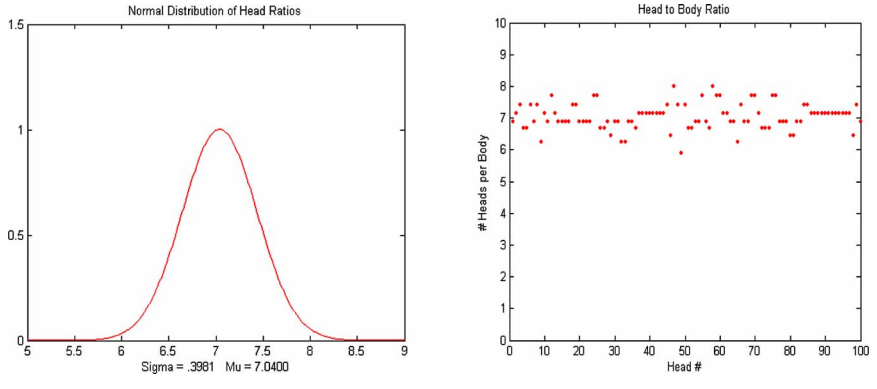


Fig. 8. Normal distribution of heads per body (left) and spread of actual number of heads per body size for all models (right)

7.3 Privacy Algorithm

What is privacy? Privacy is vulnerability. People are concerned about their own or spouses' bodies for sexual reasons. Men or women would feel insecure to exposing their private parts because of the fear of imperfectness. In addition, men feel insecure as spouses or girlfriends expose their body to other men. Overall, religion and culture play a great role here.

The rapidly growing market for 3D holographic imaging systems has sparked both interest in security applications and privacy concerns for travelers. Current body scanners use a millimeter-wave transceiver to reflect the signal of the human

body-and any objects it carries, including weapons hidden under clothing. However, these high-resolution scanned images also reveal human body details[2]. This has caused airport and transportation officials in several countries to refuse to test the scanners until researchers find more suitable ways to conceal certain parts of the human body.

At Carnegie Mellon University's Ambient Intelligence Lab, we are working to develop a method that can efficiently find and conceal humans' private parts [63]. Developing a system that's both robust and fast, however, presents a challenge. Many machine-learning algorithms are coordinate-dependent and the training data space limits them. Some algorithms only work within small bounding boxes, which is unacceptable since the system must detect the boxes before the algorithm is executed, and the boxes often aren't amenable to noise. For example, a feature-detection algorithm takes one hour to process, too long to be useful for a security screening system.

Fig. 9. The three-dimensional holographic imaging systems can detect contraband beneath clothing, yet they raise privacy concerns due to the detailed human figure that is revealed

The privacy-aware computer algorithm we developed uses the scanner to create a 3D point cloud around the human body. Since the millimeter-wave signal can't penetrate the skin, the scanner generates a 3D human surface. Furthermore, since subjects undergoing a security search are typically standing with their arms to the side, we can segment the 3D dataset into 2D contours, which significantly reduces the amount of data processing.

Unfortunately, examining each slice from top to bottom is computationally expensive. To solve that problem, we used analogous measurements invariant to coordinate systems to reduce the search space with intrinsic proportions, for example, using the height of a person's head to locate the chest.

[2] www.pnl.gov/energyscience/01-02/art1.htm

Furthermore, to determine whether the shape of a 2D contour contains defined features, we use coordinate-invariant shape properties such as height ratios and area ratios that are independent from particular poses or a specific coordinate system.

Template matching is an image-registration process that matches a surface containing known relevant information to a template of another surface. A similarity function matches the two surfaces.

Two issues arise in applying template matching on the regions of interest. First, you need to create a suitable template. Second, you must select a similarity function so that a minimization algorithm can align the template onto the region of interest. For each plane of the scan data, you can remove the back part of the body contour. By assigning the horizontal axis between the two points with the greatest distance, the model can obtain the front part of the body contour. We used radial basis functions (RBF) to configure the template for a female breast pattern, for example, and the nonlinear regression mode to match the template with the scan data.

We tested the algorithm with a database subset from the Civilian American and European Surface Anthropometry Resource project[3]. The subset contained data for 50 males and 50 females age 16 to 65. Fifty of the subjects were North American, 24 were Asian, and 26 were from Italy and the Netherlands.

We designed the experiment to test whether the algorithm can find the breast features from known female and male scan data samples. Figure 2 shows these test results.

From the plot, we can see that two distinguishable groups coincide with each subject's gender. The male subjects tend to have no curvature features and lie in the lower-left range of the graph, whereas female subjects demonstrate curvature features and lie in the upper-right range of the graph. There's a "dilemma" zone where some overweight males do have curvature features. However, the overlapped zone is small, less than 8 percent of the total 100 samples.

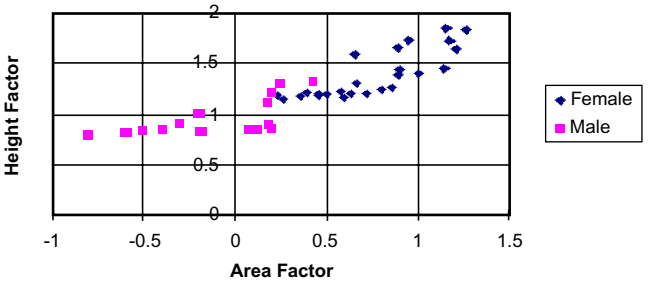


Fig. 10. Classification results. Template matching separated males without curvature features from females with curvature features.

After we calculate the area and height factors, we determine the private feature areas. Once the system finds those areas, it reduces the polygon resolution so that the area is either blurred or transparent. Fig. 10 and 12 show the results of blurring and transparency, respectively.

[3] CAESAR database: www.sae.org/technicalcommittees/caesar.htm

To determine the usefulness of these techniques in meeting privacy concerns, we conducted empirical usability tests based on two sets of the surface-rendering methods: blurring or transparency at six levels, shown as Fig. 10 and 11. The study included 10 males and 10 females, age 19 to 57.

7.4 Priority of Instincts

We have conducted two experiments that reveal how people prioritize instincts between security and privacy [63]. As we know, security has higher priority over privacy under a certain circumstances.

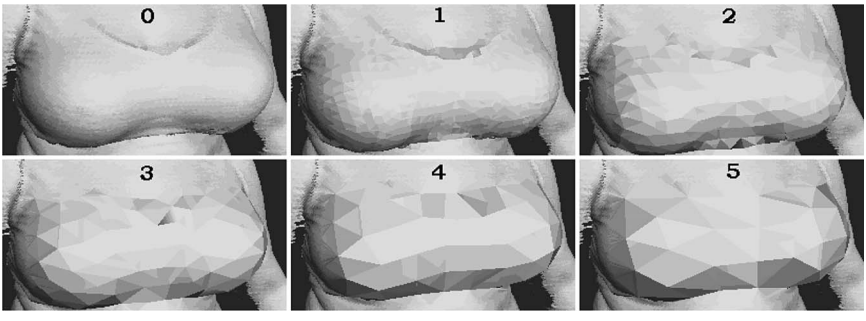


Fig. 11. The blurring scale

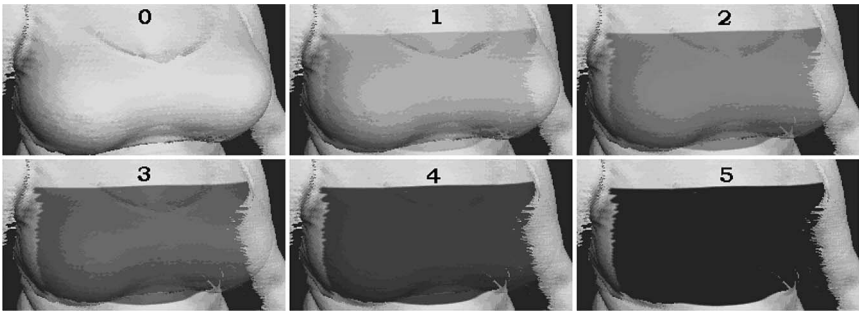


Fig. 12. The transparent scale

In the first study, we told the male subjects to imagine they (or their girlfriends or wives) had to walk through a 3D holographic scanner that would display resulting images to airport security officials on duty. They were asked to choose between a blurred or transparent image. The men averaged a 4.8 on the blurred scale and a 4.2 on the transparent scale. When asked about their concern regarding walking through the scanner, the women averaged a 4.0 on the blurred scale and a 3.8 on the transparent scale.

In the second study, we told subjects to rate their concern about privacy versus security in a scenario where we would also observe others possibly trying to conceal

weapons. Such oddities as a pocketknife between the breasts would be more difficult to detect in a very blurred mesh. The men averaged a 3.2 on the blurred scale and a 2.9 on the transparent scale. The women, on the other hand, averaged a 2.5 on the blurred scale and a 2.3 on the transparent scale.

The two usability studies indicate that different contexts can affect a subject's response and personal choice. In the first study, the men were more concerned about having their girlfriends or wives seen than the women were with how much they were seen. In the second study, almost every subject was willing to give up more privacy for the benefits of security and travel safety.

The development of the privacy algorithm is preliminary, and considers only one privacy-concerned area. However, it offers a feasible way to develop more robust privacy algorithms with available anthromorphics and physics databases. The method has been tested on 100 datasets from the CAESAR database. In both the blurring and transparency surface-rendering methods we studied, test subjects preferred to have the most privacy possible. However, the subjects adjusted their privacy concerns to a certain degree as they were informed of the security context.

8 Learning by Virtual Experiencing

According to Baldwin Effect [39], we are able to update our instincts from persistent learning. The result is automation. Many commonsense operations can be executed subconsciously. This principle can be applied to both human and machine. Fortunately, virtual reality technologies enable us to experience a simulated world without fatal risks, such as illness, falling and death.

8.1 BioSim Game [64]

According to Constructism, we do not simply learn things but construct them. Understanding and solving biomedical problems requires insight into the complex interactions between the components of biomedical systems by domain and non-domain experts. This is challenging because of the enormous amount of data and knowledge in this domain. Therefore, non-traditional educational tools have been developed such as a biological storytelling system, animations of biomedical processes and concepts, and interactive virtual laboratories.

We developed a computer game to allow children to explore biomedical knowledge [64]. We designed a biological world model, in which users can explore biological interactions by role-playing “characters” such as cells and molecules or as an observer in a “shielded vessel”, both with the option of networked collaboration between simultaneous users. The system architecture of these “characters” contains four main components: 1) bio-behavior is modeled using cellular automata, 2) bio-morphing uses vision-based shape tracking techniques to learn from recordings of real biological dynamics, 3) bio-sensing is based on molecular principles of recognition to identify objects, environmental conditions and progression in a process,

4) bio-dynamics implements mathematical models of cell growth and fluid-dynamic properties of biological solutions.

The principles are implemented in a simple world model of the human vascular system and a biomedical problem that involves an infection by Neisseria meningitides where the biological characters are both white and red blood cells and Neisseria cells. Our case studies show that the problem-solving environment can inspire user's strategic and creative thinking. Fig. 12 and 13 shows an example of the learning process.

8.2 Virtual Reality Training for Robots

The feasibility of using virtual reality to train nonhumans is demonstrated in studies that used virtual reality with rats. Hölscher et al [65] found that rats are capable of navigating in virtual environments. In parallel, Nekovarova & Klement [66] found that rats can learn to press levers for rewards when certain configurations appear on a computer screen. Therefore, while previously only humans and primates were thought to be able to navigate virtual worlds, these rat studies show that non-primates can be trained to navigate them as well.

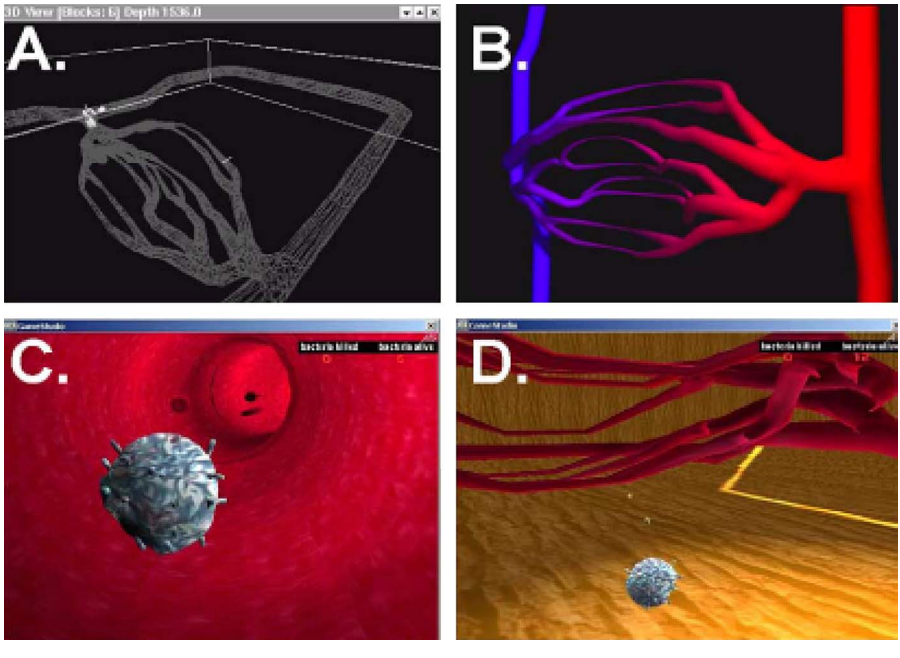


Fig. 13. World model and biological characters implemented in the game. (A) Wireframe model of the vascular system world model; (B) 3D photorealistic model showing arteries and capillaries; (C) the macrophage (white blood cell) is inside the blood stream with red blood cells; (D) after actively moving out of the blood stream, the macrophage approaches bacteria that infected human body tissue.

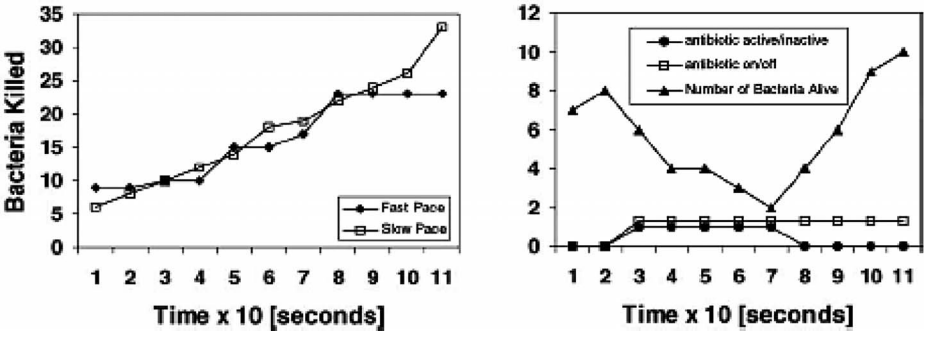


Fig. 14. Strategies against bacterial infection that can be explored in the game, speed of macrophage movement (left) and use of antibiotics (right). The fast pace strategy often leads to missing targets. The slow pace strategy gains steady capture rate. The use of antibiotics can be limited by gradual development of bacterial resistance. At that point, administration of drug does not inhibit bacterial growth.

If rats can navigate in virtual reality and humans can train in virtual reality, can robots be trained in virtual reality? If so, because humans who are trained in virtual reality can perform equal to or better than humans who are trained in real life, the possibility exists that training a robot in virtual reality may result in improved (faster and more accurate) performance than training a robot in real life.

Fig. 15. "How many robots to go before they can learn death?" The death instinct is a paradox of instinctive computing. It is possible to learn it through analogy [94-95] or virtual reality. Copyright© Yang Cai, 2007.

For example, Virtual Reality Medical Center[4] developed a simulator DARWARS [67] that enables soldiers to navigate a virtual shoot house before running through a real-life shoot house that had a different layout; and their improved speed in the shoot

[4] Dr. Mark Wiederhold, president of VRMC: www.vrphobia.com is a principal investigator of the project.

house demonstrates that spatial skills learned in virtual reality transferred to real life. There is thus reason to believe that if a robot were trained in navigational skills in virtual reality, it would learn improved spatial skills that transfer to previously unexplored environments.

On the other hand, *virtual reality enables robots to learn instinctive computing* that are impossible or too expensive to learn in real-life, such as learning the concept of 'death.'

9 Ambient Intelligence

As the volume and complexity of information grows exponentially, information-overload becomes a common problem in our life. We find an increasing demand for intelligent systems to navigate databases, spot anomalies, and extract patterns from seemingly disconnected numbers, words, images and voices. Ambient Intelligence (AmI) is an emerging paradigm for knowledge discovery, which originally emerged as a design language for invisible computing [68] and smart environment [13,69,70,71]. Since its introduction in the late 1990's, this vision has matured, having become quite influential in the development of new concepts for information processing as well as multi-disciplinary fields including computer science, interaction design, mobile computing and cognitive science.

As a part of Artificial Intelligence, Ambient Intelligence is a subconscious approach for ubiquitous computing, which inspires new theories and architectures for *'deep interactions'* that link to our instinct, such as empathic computing [72].

In a broad sense, Ambient Intelligence is perceptual interaction, involving common sense, serendipity, analogy, insight, sensory fusion, anticipation, aesthetics and emotion, all modalities that we take for granted.

True Ambient Intelligence requires instinct computing! We discover knowledge through the windows of our senses: sight, sound, smell, taste and touch, which not only describe the nature of our physical reality but also connect us to it. Our knowledge is shaped by the fusion of multidimensional information sources: shape, color, time, distance, direction, balance, speed, force, similarity, likelihood, intent and truth. Ambient Intelligence is not only perception but also interaction. We do not simply acquire knowledge but rather construct it with hypotheses and feedback. Many difficult discovery problems become solvable through interaction with perceptual interfaces that enhance human strengths and compensate for human weaknesses to extend discovery capabilities. For example, people are much better than machines at detecting patterns in a visual scene, while machines are better at detecting errors in streams of numbers.

10 Empathic Computing

Empathic computing aims to enable a computer to understand human states and feelings and to share the information across networks. Instinctive computing is a necessary foundation of empathic computing.

For decades, computers have been viewed as apathetic machines that only accept or reject instructions. Whether an artifact can understand human's feeling or state is a paradox of empathy. René Descartes claims that thoughts, feelings, and experience are private and it is impossible for a machine to adequately understand or know the exact feelings of people. On the other hand, Ludwig Wittgenstein states that there is no way to prove that it is impossible to adequately imagine other people's feeling [73]. Alan Turing argues that machine intelligence can be tested by dialogs through a computer keyboard [74-76]. In our case, the Turing Test can be simplified as a *time-sharing test*, where empathic machines and humans coexist in a care-giving system with a time-sharing schedule. If a person receives care continuously, then we may call the system 'empathic'.

Empathic computing emerges as a new paradigm that enables machines to know who, what, where, when and why, so that the machines can anticipate and respond to our needs gracefully. Empathic computing in this study is narrowed down to understand the 'low-level' subconscious feelings, such as pain, illness, depression or anomaly. Empathic computing is a combination of Artificial Intelligence (AI), network communication and human-computer interaction (HCI) within a practical context such as healthcare.

The AI program ELIZA is perhaps the first artifact that is capable to engage in an empathic conversation [80]. Based on simple keyword matching, the program appears to be a 'good listener' to psychiatric patients. This shows that a small program could generate pseudo-empathy to a certain degree. However, human feelings and states are more than just verbal communication. We watch, listen, taste, smell, touch and search. Warwick's project Cyborg [78] is probably the most daring physical empathic artifact. The pioneer implanted an electrode array under his skin that interfaced directly into the nervous system. The signal was fed into a robot arm that mimicked the dynamics of Warwick's own arm. Furthermore, the researcher implanted a sensor array into his wife's arm with the goal of creating a form of telepathy or empathy using the Internet to communicate the signal remotely.

Empathic sensor webs provide new opportunities to detect anomalous events and gather vital information in daily life. Their widespread availability and affordability makes it easier and cheaper to link already deployed sensors such as video cameras. New sensor web capabilities can have a major impact by changing how information is used in homecare. For example, smart mirror for tongue inspection, smart room [99] and wearable sensors for motion pattern analysis, etc.

Empathic computing brings a new paradigm to the network-centric computing, which focuses on sensor fusion and human-computer interaction. To avoid the potential information avalanches in the empathic sensor, there are instinctive computing solutions for information reduction on the source side, for instance, applying the power law for multi-resolution channel design and interacting mobile sensors with stationary sensors.

There are a few prototypes of empathic computing. For example, Fig. 15 shows the wearable empathic computing system that is to detect an instance of an individual falling down. It is not a complete system. Rather, it only shows how complicated an empathic computing could be involved. From the initial results, it is found that the location of the wearable sensor makes a difference. The belt, for example, is probably the most appropriate place to put the sensor for detecting a fall. From the machine learning algorithm, the accuracy reaches up to 90% from 21 simulated trials.

Fig. 16. Wearable sensor that detects anomalous events

With the growing need for home health care, empathic computing attracts attention from many fields. Recent studies include designing a home for elderly people or people with disabilities [79]. Healthcare systems are looking for an easy and cost-effective way to collect and transmit data from a patient's home. For example, a study [80] shows that the GSM wireless network used by most major cell phone companies was the best for sending data to hospitals from a patient's home. Universities and corporations have launched labs to explore the healthy living environment, such as LiveNet [81-82], HomeNet [83], and Philips' HomeLab [84]. Furthermore, Bodymedia has developed the armband wearable sensor [85-86] that tracks body temperature, galvanic skin response, heat flux, and other data. However, most of products have not reached the scale of economy.

11 Conclusions

Early Artificial Intelligence scholars used a 'top-down' approach to study human behavior. They focused on logic reasoning and languages. The more they went deep into the human's mind, the more they felt the need to incorporate human instincts including perceptual intelligence, empathy, and commonsense. For example, Herb Simon combined logic model with pictorial representation in CaMeRa model [87]. John Anderson extended ACT-R [27] with sensory function. However, the 'top-down' approaches have their limitations. Herb Simon saw the gap between the sequential logic reasoning and parallel sensing. "Have you found any neural network that is capable of solving the Hanoi Tower Problem?" He challenged his students. Perhaps instinctual computing is a potential bridge between the two.

The fundamental difference between existing machines and a living creature is *Instinct!* Instinctive computing is a computational simulation of biological and cognitive instincts. It is a meta-program of life, just like universal gravity in nature. It profoundly influences how we look, feel, think, and act. If we want a computer to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even *to have* and express primitive instincts.

In this paper, we reviewed the recent work in this area, the instinctive operating system, and potential applications. The paper proposes a 'bottom-up' approach that is to focus on human basic instincts: forage, vigilance, reproduction, intuition and learning. They are the machine code in human operating systems, where high-level programs, such as social functions can override the low-level instincts. However, instinctive computing has been always a default operation. Instinctive computing is the foundation for Ambient Intelligence as well as Empathic Computing.

In his book "Philosophical Investigations", Ludwig Wittgenstein states: "The aspects of things that are most important for us are hidden because of their simplicity and familiarity. One is unable to notice something - because it is always before one's eyes. The real foundations of his enquiry do not strike a man at all. Unless that fact has at some time struck him. And this means: we fail to be struck by what, once seen, is most striking and most powerful." Perhaps, Instinctive Computing is a key to unveiling the hidden power in human dynamics.

Acknowledgement

The author would like to thank Brian Zeleznik, Daniel Sonntag and Maja Pantic for their insightful comments. Thanks to my research assistants Guillaume Milcent, Russell Savage, and Deena Zytnick at Carnegie Mellon University.

References

1. Freud S. Instincts and their Vicissitudes, 1915, Psychoanalytic Electronic Publishing, http://www.p-e-p.org/pepcd.htm
2. Albrecht-Buehler, G. Is Cytoplasm Intelligent too? In: Muscle and Cell Motility VI (ed. J. Shay) p. 1-21, 1985
3. Albrecht-Buehler, G., http://www.basic.northwestern.edu/g-buehler/cellint0.htm, updated 21 Jan. 2007
4. von Neumann, J. (1966), the Theory of self reproducing automata. Edited by A. Burks, Univ. of Illinois Press, Urbana.
5. Conway, J. Game of Life, Wikipedia: http://en.wikipedia.org/wiki/Conway%27s_Game_of_Life
6. BioWall: http://lslwww.epfl.ch/biowall/
7. Wolfram, S. The new kind of science, Wolfram Media, 2000
8. Jirí Kroc: Model of Mechanical Interaction of Mesenchyme and Epithelium in Living Tissues. 847-854, in Vassil N. Alexandrov, G. Dick van Albada, Peter M. A. Sloot, Jack Dongarra (Eds.): Computational Science - ICCS 2006, 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings, Part IV. Lecture Notes in Computer Science 3994 Springer 2006, ISBN 3-540-34385-7
9. Regirer, S.A. and Shapovalov, D.S. Filling space in public transport by passengers, Automation and Remote Control, vol. 64, issue 8, August 2003
10. Xilinx, www.xilinx.com, captured in 2007
11. MacKinnon, N. Symbolic interaction as affect control, State University of New York Press, 1994
12. Mueller, E. T. 1990. Daydreaming in Humans and Machines: A computer model of the stream of thought. Norwood, NJ

13. Pentland, A. 1996. Perceptual intelligence, in Bowyer, K. and Ahuja, N. (eds) “Image Uniderstanding, IEEE Computer Society Press, 1996
14. Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., Shepard, B. 2006. From Cyc to Intelligent Assistant, in Cai, Y. and Abascal, J. (eds), Ambient Intelligence for Everyday Life, LNAI 3864, Springer, 2006
15. Picard, R. Affective computing, The MIT Press, 1998
16. Minsky, M. The emotion machine, Simon and Schuster, 2006
17. Cohen, H. 1995. The Further Exploits of AARON, Painter, Stanford Electronic Humanities Review. Vol. 4, No. 2.
18. Leyton, M. Symmetry, causality, mind, The MIT Press, 1999
19. Guare, J. Six Degrees of Separation, Vintage, 1990
20. Albert-Laszlo Barabasi, Linked: The New Science of Networks, Perseus, 2002
21. Eagle, N. and A. Pentland (2005), "Reality Mining: Sensing Complex Social Systems", *Personal and Ubiquitous Computing*, September 2005
22. Chakrabarti, D., Y. Zhan, D. Blandford, C. Faloutsos and G. Blelloch, NetMine: New Mining Tools for Large Graphs, in the SDM 2004 Workshop on Link Analysis, Counter-terrorism and Privacy
23. Kim Rossmo, Geograpical Profiling, CRC Press, 1990, ISBN: 0849381290
24. Helbing, D., Illés Farkas and Tamás Vicsek, Simulating dynamical features of escape panic, *Nature* 407, 487-490, number 28 September 2000
25. Angell,L.S., Young, R.A., Hankey, J.M. and Dingus, T.A. , 2002, An Evaluation of Alternative Methods for Assessing Driver Workload in the Early Development of In-Vehicle Information Systems, SAE Proceedings, 2002-01-1981
26. Salvucci, D. D. (2005). Modeling tools for predicting driver distraction. In Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting. Santa Monica, CA: Human Factors and Ergonomics Society.
27. Anderson, J. ACT-R, http://act-r.psy.cmu.edu/
28. Bonabeau, E., Dorigo, M. and Theraulaz, G. Swarm Intelligence: from natural to artificial systems, Oxford University Press, 1999
29. Smith, R.: Alarm signals in fishes. Rev Fish Biol Fish 2:33-63, 1992
30. McClintock, M.K. (1984). Estrous synchrony: modulation of ovarian cycle length by female pheromones. Physiological Behavior 32, 701-705
31. Wyatt, Tristram D. (2003). Pheromones and Animal Behaviour: Communication by Smell and Taste. Cambridge: Cambridge University Press. ISBN 0521485266.
32. BBC program: Human instinct: http://www.bbc.co.uk/science/humanbody/tv/humaninstinct/
33. Edwards, B. Drawing on the right side of brain, Jeremy P. Tarcher / Putnam, 1999
34. Li, Z. and Guyader, N. Interference with Bottom-Up Feature Detection by Higher-Level Object Recognition, Current Biology 17, 26–31, January 9, 2007, Elsevier Ltd
35. Robertsson, L., Iliev, B., Palm, R., Wide, P. 2005. Perception Modeling for Human-Like Artificial Sensor Systems. International Journal of Human-Computer Studies, V. 65, No.5, 2007
36. Krepki, R., Miller, K.R., Curio, G. and Balnkertz, B. 2006. Brain-Computer Interface - An HCI-Channel for Discovery, this issue
37. von Nuemann cellular automata, Wikipedia, http://en.wikipedia.org/wiki/Von_Neumann_cellular_automata,
38. Forsyth D.A.; Fleck, M.M., Identifying nude pictures, *Proceeding. Third IEEE Workshop on Applications of Computer Vision.* 103-108, 1996.

39. Baldwin, Mark J. A New Factor in Evolution. The American Naturalist, Vol. **30**, No. 354 (Jun., 1896), 441-451.
40. Mueller, E. T. 1990. Daydreaming in Humans and Machines: A computer model of the stream of thought. Norwood, NJ: Ablex
41. Mueller, E. T. 2000. A Calendar with Common Sense. Proceedings of the 2000 International Conference on Intelligent User Interfaces (pp. 198-201). New York: Association for Computing Machinery.
42. Mueller, E. T. 2003. Story Understanding through Multi-Representation Model Construction. In Graeme Hirst & Sergei Nirenburg (Eds.), Text Meaning: Proceedings of the HLT-NAACL 2003 Workshop (pp. 46-53). East Stroudsburg, PA: Association for Computational Linguistics
43. Beale, R. 2006. Supporting Serendipity in Data Mining and Information Foraging, IJCHS, vol. 65, num. 5, 2007
44. Roberts, R. 1989. Serendipity – Accidental Discoveries in Science, John Wiley & Sons, Inc.
45. Wong, P.C. 1999. Visual Data Mining, IEEE Visual Data Mining. IEEE Computer Graphics and Applications, Vol. 19, No. 5, Sept. 1999
46. Tanz, O. and Shaffer, J. Wireless local area network positioning, in Cai, Y. (ed) Ambient Intelligence for Scientific Discovery, LNAI 3345, Springer, 2005
47. Smailagic, A., Siewiorek, D. P., Anhalt, J., Kogan, D.,Wang, Y.: "Location Sensingand Privacy in a Context Aware Computing Environment." Pervasive Computing. 2001
48. Georges G. Grinstein Usama Fayyad and Andreas Wierse, editors. Information Visualization in Data Mining and Knowledge Discovery, chapter 2, pages 58–61. 2001.
49. S. Lesteven P. Poinot and F. Murtagh. "A spatial user interface to the astronomical literature." aaps, 130:183–191, may 1998
50. Java tools for experimental mathematics. http://www-sfb288._math.tu-berlin.de/˜jtem/numericalMethods/download.html , 2004.
51. Simplex Optimization: http://www.grabitech.se/algorithm.htm
52. Levin, I. "KDD-99 Classifier Learning Contest: LLSoft's Results Overview", ACM SIGKDD Explorations2000, pp. 67-75, January 2000.
53. Milcent, G. and Cai, Y. Flow-On-Demand for network traffic control, Proceedings of Ambient Intelligence for Life, Spain, 2005
54. Zhang, E.: Diagnostics of Traditional Chinese Medicine. Publishing House of Shanghai University of Traditional Chinese Medicine, ISBN 7-81010-125-0. in both Chinese and English. (1990)
55. Gunarathne, G.P. Presmasiri, Gunarathne, Tharaka R.: Arterial Blood-Volume Pulse Analyser. IEEE, Instrumentation and Measurement Technology Conference, AK, USA, May. (2002) 1249-1254
56. Schnorrenberger, C. and Schnorrenberger, B. Pocket Atlas of Tongue Diagnosis, Thieme, Stuttgart, New York, 2005
57. Cai, Y et al, : Ambient Diagnostics, in Y. Cai (ed.) Ambient Intelligence for Scientific Discovery, LNAI 3345, pp. 248-262, 2005
58. Singh, D. Renn, P. and Singh, A. Did the perils of abdominal obesity affect depiction of feminine beauty in the sixteenth to eighteenth century British literature? Exploring the health and beauty link, Proceedings of the Royal Society B: Biological Sciences, Vol. 274, No. 1611, March 22, 2007
59. Suikerbuik C.A.M. Automatic Feature Detection in 3D Human Body Scans. Master thesis INF/SCR-02-23, Institute of Information and Computer Sciences. Utrecht University, 2002

60. Suikerbuik R., H. Tangelder, H. Daanen, A. Oudenhuijzen, Automatic feature detection in 3D human body scans, Proceedings of SAE Digital Human Modeling Conference, 2004, 04-DHM-52
61. Goldgof D.B., T. S. Huang, and H. Lee, "Curvature based approach to terrain recognition," Coord. Sci. Lab., Univ. Illinois, Urbana-Champaign, Tech. Note ISP-910, Apr. 1989.
62. Fleck, M.M., D.A. Forsyth and C. Bregler, Finding naked people, *Proc. European Conf. on Computer Vision* , Edited by: Buxton, B.; Cipolla, R. Berlin, Germany: Springer-Verlag, 1996. p. 593-602
63. Laws, J., N. Bauernfeind and Y. Cai, Feature Hiding in 3D Human Body Scans, Journal of Information Visualization, Vol.5, No. 4, 2006
64. Cai, Y., Snel, I., Bharathi, B.S., Klein, C. and Klein-Seetharaman, J. 2003. Towards Biomedical Problem Solving in a Game Environment. In: Lecture Notes in Computer Science 2659, 1005-1014.
65. Hölscher et al , Rats are able to navigate in virtual environment, Journal of Experimental Biology, 208, 561-569 (2005): http://jeb.biologists.org/cgi/content/full/208/3/561
66. Nekovarova, T. and Klement, D. Operant behavior of the rat can be controlled by the configuration of objects in an animated scene displayed on a computer screen, Physiological research (Physiol. res.) **ISSN** 0862-8408, 2006, vol. 55, num. 1, pp. 105-113
67. DARWARS: http://www.darwars.bbn.com/
68. Norman, D. 2003. The Invisible Computer, The MIT Press
69. Aarts, E. and Marzano, S. 2004. The New Everyday – Views on Ambient Intelligence, ISBN 90 6450 5020
70. Cai, Y. 2005. Ambient Intelligence for Scientific Discovery, Lecture Notes in Artificial Intelligence, LNAI 3345, Springer, 2005
71. Cai, Y. and Abascal, J. Ambient Intelligence for Everyday Life, Lecture Notes in Artificial Intelligence, LNAI 3864, Springer, 2006
72. Cai, Y. Empathic Computing, in Cai, Y. and Abascal, J. (eds), Ambient Intelligence for Everyday Life, Lecture Notes in Artificial Intelligence, LNAI 3864, Springer, 2006
73. Moore, G.: Cramming more components onto integrated circuits. Electronics, Vol. 38, No. 8, April 19. (1965)
74. Popple, A.V.: The Turing Test as a Scientific Experiment. Psychology, 7(15), 1996
75. Turing, A.M.: Computing Machinery and Intelligence. Mind, 59: 433-460, 1950
76. Weizenbaum, J. Computer Power and Human Reason: From Judgment To Calculation, San Francisco: W. H. Freeman, 1976 ISBN 0-7167-0463-1
77. Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine, Communications of the Association for Computing Machinery 9 (1966): 36-45.
78. Williams, J. A., Dawood, A. S. and Visser, S. J.: FPGA-Based Cloud Detection for Real-Time Onboard Remote Sensing, IEEE ICFPT 2002, Hong Kong
79. Dewsburry, Guy, Taylor, Bruce, Edge, Martin: Designing Safe Smart Home Systems for Vulnerable People.
80. Herzog, A., and Lind, L.: Network Solutions for Home Health Care Applications. Linkoping University (2003)
81. Sung, M. and A. Pentland, MIThril LiveNet: Health and Lifestyle Networking, Workshop on Applications of Mobile Embedded Systems (WAMES'04) at Mobisys'04, Boston, MA, June, 2004
82. LiveNet: http://hd.media.mit.edu/livenet/

83. HomeNet:http://homenet.hcii.cs.cmu.edu/, http://www.coe.berkeley.edu/labnotes/1101smartbuildings.html http://www7.nationalacademies.org/cstb/wp_digitaldivide.pdf
84. HomeLab: http://www.research.philips.com/technologies/misc/homelab/
85. Bodymedia: www.bodymedia.com
86. Farringdon, Jonathan, and Sarah Nashold: Continuous Body Monitoring. Ambient Intelligence for Scientific Discovery (2005): 202-223.
87. Tabachneck-Schijf H.J.M., Leonardo, A.M. , and Simon. CaMeRa : A computational mode of multiple representations, Journal of Cognitive Science, vol. 21, num.3, pp. 305-350, 1997: http://cat.inist.fr/?aModele=afficheN&cpsidt=2108318
88. Anthropometry Resource (CAESAR), Final Report, Volume I: Summary, AFRL-HE-WP-TR-2002-0169, United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, 2255 H Street, Wright-Patterson AFB OH 45433-7022 and SAE International, 400 Commonwealth Dr., Warrendale, PA 15096.
89. Maslow, A. H. (1943). A Theory of Human Motivation. Psychological Review, 50, 370-396.
90. Simon, H. The science of the artificial, 3rd edition, The MIT Press, 1996
91. Simon, H.A.1989, Models of Thought, Volume II, Yale University Press, 1989
92. Midgley, M. Beast and Man: the roots of human nature, The Harvester Press, 1979
93. Köhler, W. 1947. Gestalt Psychology, Liveright, New York
94. Mitchell, M. 1993. Analogy-Making as Perception: A Computer Model, The MIT Press
95. Holyoak, K. and Thagard, P. 1994. Mental Leaps: Analogy in Creative Thought, The MIT Press
96. Rosen, R. Anticipatory Systems, Pergamon Press, 1985
97. Wiener, N. Cybernetics: or control and communication in the animal and the machine, The MIT Press, 1961
98. Boff KR and Kaufman, L, and Thomas, JP (eds), Human Performance Measures Handbook, Wiley and Sons, 1986
99. Pentland, A. “Smart Rooms, Smart Clothes, Scientific American, June, 1996
100. Instinct theory, Wikipedia, http://en.wikipedia.org/wiki/Instinct_theory.
101. King, B.M. Human sexuality today, 3rd edition, Prentice-Hall International, 1991
102. Lewin, K. A dynamic theory of personality. New York: McGraw-Hill, 1935
103. Lewin, K. The conceptual representation and measurement of psychological forces. Contr. psychol. Theor., 1938, 1(4).
104. Pantic, M., A. Pentland, A. Nijholt and T.S. Huang, Human Computing and machine understanding of human behavior: A Survey, Proc. ACM Int'l Conf. Multimodal Interfaces, 2006, pp. 239-248

Human Computing and Machine Understanding of Human Behavior: A Survey

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

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

Abstract. A widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living spaces and projecting the human user into the foreground. If this prediction is to come true, then next generation computing should be about anticipatory user interfaces that should be human-centered, built for humans based on human models. They should transcend the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating certain human behaviors such as affecti0ve and social signaling. This article discusses how far are we from enabling computers to understand human behavior.

Keywords: Human sensing, Human Behavior Understanding, Multimodal Data Analysis, Affective Computing, Socially-aware Computing.

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, emphasizing 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, emphasizing once again physical environments. Futuristic movies often contain such visions of human environments of the future – fitted out with arrays of intelligent, yet invisible devices, homes, transportation means and working spaces of the future can anticipate every need of their inhabitants (Fig. 1). In this vision of the future, often referred to as "ubiquitous computing" [87] or "ambient intelligence" [1], 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 favorite background music will play when we come home showing signs of weariness.

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 47–71, 2007.

Fig. 1. Human-centered environments of the future envisioned in SF movies (left to right): intelligent robotic assistants (*Artificial Intelligence*, 2001), smart operation chamber (*Firefly*, 2002), speech-driven and iris-ID-based car (*Minority Report*, 2002)

Although profoundly appealing, this vision of the digital future creates a set of novel, greatly challenging issues [48], [77], [90], [49], [50]. It assumes a shift in computing – from desktop computers to a multiplicity of smart computing devices diffused into our environment. It assumes that computing will move to the background, weave itself into the fabric of everyday living spaces and disappear from the foreground, projecting the human user into it. However, as computing devices disappear from the scene, become invisible, weaved into our environment, how the interaction between this technology and humans will evolve? 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 the key, how do we arrive at context-aware systems?

Human computer interaction (HCI) designs were first dominated by direct manipulation and then delegation. Both styles of interaction involve usually the conventional interface devices like keyboard, mouse, and visual displays, and assume that the human will be explicit, unambiguous and fully attentive while controlling information and command flow. This kind of interfacing and categorical computing works well for context-independent tasks like making plane reservations and buying and selling stocks. However, it is utterly inappropriate for interacting with each of the (possibly hundreds) computer systems diffused throughout future smart environments and aimed at improving the quality of life by anticipating the users needs. Clearly, "business as usual" will not work in this case. We must approach HCI in a different way, moving away from computer-centered designs toward human-centered designs for HCI, made for humans based on models of human behavior. Human-centered designs will require explorations of *what* is communicated (linguistic message, nonlinguistic conversational signal, emotion, attitude), *how* the information is passed on (the person's facial expression, head movement, nonlinguistic vocalization, hand and body gesture), *why*, that is, in which context the information is passed on (where the user is, what his or her current task is, are other people involved), and *which* (re)action should be taken to satisfy user needs and requirements. Modeling human behavior is a challenging task, however. How far are we from attaining it?

2 Issues in Modeling Human Behavior

Instead of focusing on the computer portion of the HCI context, designs for human-centered computing should focus on the human portion of the HCI context. They should go beyond the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating certain human behaviors like affective and social signaling. The design of these functions will require explorations of the following.

- What is communicated? Which type of message is communicated by shown behavioral signals (linguistic message, nonlinguistic conversational signal, emotion, attitude, mood)?
- How the information is passed on (the person's facial expression, head movement, hand and body gesture, nonlinguistic vocalization)? This question is closely related to issues such as which human communicative cues convey information about human behaviors like social and emotional signaling and which modalities should be included into an automatic analyzer of human behavioral signals.
- Why, that is, in which context the information is passed on (where the user is, what his or her current task is, are other people involved)? This question is related to issues such as what to take into account to realize interpretations of shown behavioral signals (e.g., the person's identity, current task), how to distinguish between different types of messages (e.g., emotions vs. conversational signals), and how best to integrate information across modalities given the context in which the information is passed on.

2.1 What Is Communicated?

The term behavioral signal is usually used to describe a set of temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (a blink) to minutes (talking) or hours (sitting). Among the types of messages conveyed by behavioral signals are the following [24] (Fig. 2):

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

While there is agreement across different theories that at least some behavioral signals evolved to communicate information, there is lack of consensus regarding their specificity, extent of their innateness and universality, and whether they convey emotions, social motives, behavioral intentions, or all three [38]. Arguably the most often debated issue is whether affective states are a separate type of messages communicated by behavioral signals (i.e. whether behavioral signals communicate

Fig. 2. Types of messages conveyed by behavioural signals. First row: affective states (anger, surprise, disbelief, sadness). Second row: emblems (wink, thumbs up), illustrators and regulators (head tilt, jaw drop, look exchange, smile), manipulators (yawn).

actually felt affect), or is the related behavioral signal (e.g. facial expression) just an illustrator / regulator aimed at controlling "the trajectory of a given social interaction", as suggested by Fridlund [28]. Explanations of human behavioral signals in terms of internal states such as affective states are typical to psychological stream of thought, in particular to discrete emotion theorists who propose the existence of six or more basic emotions (happiness, anger, sadness, surprise, disgust, and fear) that are universally displayed and recognized from non-verbal behavioral signals (especially facial and vocal expression) [43], [40]. Instead of explanations of human behavioral signals in terms of internal states, ethologists focus on consequences of behavioral displays for interpersonal interaction. As an extreme within the ethological line of thought, social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. According to Fridlund, facial expressions should not be labeled in terms of emotions but in terms of Behavioral Ecology interpretations, which explain the influence a certain expression has in a particular context [28]. Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. However, as proposed by Izard [38], one may feel angry without the slightest intention of attacking anyone. In summary, is social communication the sole function of behavioral signals? Do they never represent visible manifestation of emotion / feeling / affective states? Since in some instances (e.g. arachnophobia, acrophobia, object-elicited disgust, depression), affective states are not social, and their expressions necessarily have aspects other than "social motivation", we believe that affective states should be included into the list of types of messages communicated by behavioral signals. However, it is not only discrete emotions like surprise or anger that represent the affective states conveyed by human behavioral signals. Behavioral cues identifying attitudinal states like interest and

boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are essential components of human behavior. Hence, in contrast to traditional approach, which lists only (basic) emotions as the first type of message conveyed by behavioral signals [24], we treat affective states as being correlated not only to emotions but to other, aforementioned social signals and attitudinal states as well.

2.2 How the Information Is Passed on?

Manipulators are usually associated with self-manipulative gestures like scratching or lip biting and involve facial expressions and body gestures human communicative cues. Emblems, illustrators and regulators are typical social signals, spoken and wordless messages like head nods, bow ties, winks, 'huh' and 'yeah' utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech. The most complex messages communicated by behavioral signals are affective and attitudinal states. Affective arousal modulates all human communicative signals. Hence, one could expect that automated analyzers of human behavior should include all human interactive modalities (audio, visual, and tactile) and should analyze all verbal and non-verbal interactive signals (speech, body gestures, facial and vocal expressions, and physiological reactions). However, we would like to make a few comments here.

Although spoken language is between 200 thousand and 2 million years old [31], and speech has become the indispensable means for sharing ideas, observations, and feelings, findings in basic research indicate that in contrast to spoken messages [29], nonlinguistic messages are the means to analyze and predict human behavior [2]. Anticipating a person's word choice and the associated intent is very difficult [29]: even in highly constrained situations, different people choose different words to express exactly the same thing.

As far as nonverbal cues are concerned, it seems that not all of them are equally important in the human judgment of behavioral signals. People commonly neglect physiological signals, since they cannot sense them at all times. Namely, in order to detect someone's clamminess or heart rate, the observer should be in a physical contact (touch) with the observed person. Yet, the research in psychophysiology has produced firm evidence that affective arousal has a range of somatic and physiological correlates including pupillary diameter, heart rate, skin clamminess, temperature, respiration velocity [10]. This and the recent advent of non-intrusive sensors and wearable computers, which promises less invasive physiological sensing [74], open up possibilities for including tactile modality into automatic analyzers of human behavior [62]. However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues [2]. Human judges seem to be most accurate in their judgment when they are able to observe the face and the body. Ratings that were based on the face and the body were 35% more accurate than the ratings that were based on the face alone. Yet, ratings that were based on the face alone were 30% more accurate than ratings that were based on the body alone and 35% more accurate than ratings that were based on the tone of voice alone [2]. These findings indicate that to interpret someone's behavioral cues, people rely on shown facial expressions and to a lesser degree on shown body

gestures and vocal expressions. Note, however, that gestures like (Fig. 2) scratching (manipulator), thumbs up (emblem), finger pointing (illustrator), and head nods (regulator) are typical social signals. Basic research also provides evidence that observers tend to be accurate in decoding some negative basic emotions like anger and sadness from static body postures [17] and that gestures like head inclination, face touching, and shifting posture often accompany social affective states like shame and embarrassment [16]. In addition, although cognitive scientists were unable to identify a set of vocal cues that reliably discriminate among affective and attitudinal states, listeners seem to be rather accurate in decoding some basic emotions from vocal cues like pitch and intensity [40] and some non-basic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns [67]. Thus, automated human behavior analyzers should at least include facial expression and body gestures modalities and preferably they should also include modality for perceiving nonlinguistic vocalizations.

Finally, while too much information from different channels seem to be confusing to human judges, resulting in less accurate judgments of shown behavior when three or more observation channels are available (face, body, and speech) [2], combining those multiple modalities (including physiology) may prove appropriate for realization of automatic human behavior analysis.

2.3 In Which Context Is the Information Passed on?

Behavioral signals do not usually convey exclusively one type of messages but may convey any of the types (e.g. scratching is usually a manipulator but it may be displayed in an expression of confusion). It is crucial to determine to which class of behavioral signals a shown signal belongs since this influences the interpretation of it. For instance, squinted eyes may be interpreted as sensitivity of the eyes to bright light if this action is a reflex (a manipulator), as an expression of disliking if this action has been displayed when seeing someone passing by (affective cue), or as an illustrator of friendly anger on friendly teasing if this action has been posed (in contrast to being unintentionally displayed) during a chat with a friend, to mention just a few possibilities.

To determine the class of an observed behavioral cue, one must know the context in which the observed signal was displayed. Six questions summarize the key aspects of the computer's context with respect to nearby humans:

- *Who?* (Who the observed user is? This issue is of particular importance for recognition of affective and attitudinal states since it is not probable that each of us will express a particular affective state by modulating the same behavioral signals in the same way, especially when it comes to states other than basic emotions.)
- *Where?* (Where the user is?)
- *What?* (What is the current task of the user?)
- *How?* (How the information is passed on? Which behavioral signals have been displayed?)
- *When?* (What is the timing of displayed behavioral signals with respect to changes in the environment? Are there any co-occurrences of the signals?)

- *Why?* (What may be the user's reasons to display the observed cues? Except of the user's current task/activity, the issues to be considered include the properties of the user's physical environment like lighting and noise level, and the properties of the current social situation like whether the user is alone and what is his or her affective state.)

For recognition of affective and attitudinal states it is of particular importance who the observed subject is because it is not probable that each of us will express a particular affective state by modulating the same communicative signals in the same way, especially when it comes to affective states other than basic emotions.

Since the problem of context-sensing is extremely difficult to solve (if possible at all) for a general case, we advocate that a pragmatic approach (e.g. activity/application- and user-centered approach) must be taken when learning the grammar of human expressive behavior. In addition, because of the impossibility of having users instructing the computers for each possible application, we propose that methods for unsupervised (or semi-supervised) learning must be applied. Moreover, much of human expressive behavior is unintended and unconscious; the expressive nonverbal cues can be so subtle that they are neither encoded nor decoded at an intentional, conscious level of awareness [2]. This suggests that the learning methods inspired by human unconscious problem solving processes may prove more suitable for automatic human behavior analysis than the learning methods inspired by human conscious problem solving processes [81].

Another important issue is that of multimodal fusion. A number of concepts relevant to fusion of sensory neurons in humans may be of interest [75]:

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

Thus, humans simultaneously employ the tightly coupled audio, visual, and tactile modalities. As a result, analysis of the perceived information is highly robust and flexible. Hence, one could expect that in an automated analyzer of human behavior input signals should not be considered mutually independent and should not be combined only at the end of the intended analysis, as the majority of current studies do, but that they should be processed in a joint feature space and according to a context-dependent model [59]. However, does this tight coupling persists when the modalities are used for multimodal interfaces as proposed by some researchers (e.g., [33]), or not, as suggested by others (e.g., [70])? This remains an open, highly relevant issue.

3 Can Computer Systems Understand Human Behavior?

Modeling human behavior and understanding displayed patterns of behavioral signals, involve a number of tasks.

- ♦ Sensing and analyzing displayed behavioral signals including facial expressions, body gestures, nonlinguistic vocalizations, and vocal intonations.
- ♦ Sensing the context in which observed behavioral signals were displayed.
- ♦ Understanding human behavior by translating the sensed human behavioral signals and context descriptors into a description of the shown behavior.

3.1 Human Sensing

Sensing human behavioral signals including facial expressions, body gestures, nonlinguistic vocalizations, and vocal intonations, which seem to be most important in the human judgment of behavioral cues [2], involves a number of tasks.

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

Because of its practical importance and relevance to face recognition, face detection received the most attention of the tasks mentioned above. The problem of *finding faces* should be solved regardless of clutter, occlusions, and variations in head pose and lighting conditions. The presence of non-rigid movements due to facial expression and a high degree of variability in facial size, color and texture make this problem even more difficult. Numerous techniques have been developed for face detection, i.e., identification of all regions in the scene that contain a human face [89], [44]. However, virtually all of them can detect only (near-) upright faces in (near-) frontal view. Most of these methods emphasize statistical learning techniques and use appearance features, including the real-time face detection scheme proposed by Viola and Jones [84], which is arguably the most commonly employed face detector in automatic facial expression analysis. Note, however, that one of the few methods that can deal with tilted face images represents a feature-based rather than an appearance-based approach to face detection [12].

Tracking is an essential step for human motion analysis since it provides the data for recognition of face/head/body postures and gestures. Optical flow has been widely used for head, face and facial feature tracking [85]. To address the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to occlusion, clutter, and changes in illumination, researchers in the field started to use sequential state estimation techniques like Kalman and particle filtering schemes [34]. The derivation of the Kalman filter is based on a state-space model [41], governed by two assumptions: (i) linearity of the model and (ii) Gaussianity of both the dynamic noise in the process equation and the measurement noise in the measurement

equation. Under these assumptions, derivation of the Kalman filter leads to an algorithm that propagates the mean vector and covariance matrix of the state estimation error in an iterative manner and is optimal in the Bayesian setting. To deal with the state estimation in nonlinear dynamical systems, the extended Kalman filter was proposed, which is derived through linearization of the state-space model. However, many of the state estimation problems, including human facial expression analysis, are nonlinear and quite often non-Gaussian too. Thus, if the face undergoes a sudden or rapid movement, the prediction of features positions from Kalman filtering will be significantly off. To overcome these limitations of the classical Kalman filter and its extended form in general, particle filters were proposed. The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability $p(\alpha \mid Y)$, where α is the state of a temporal event to be tracked given a set of noisy observations $Y = \{y^1, \ldots, y^-, y\}$ up to the current time instant. This means that the distribution $p(\alpha \mid Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k, then it is as if s_k was drawn from $p(\alpha \mid Y)$. By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering), particle filtering is able to track multimodal conditional probabilities $p(\alpha \mid Y)$, and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear, non-Gaussian systems. Numerous particle-filtering tracking schemes were proposed including the Condensation algorithm [37], Auxiliary Particle Filtering [63], and Particle Filtering with Factorized Likelihoods [61]. Some of the most advanced approaches to head tracking and head-pose estimation are based on Kalman (e.g., [36]) and particle filtering frameworks (e.g., [3]). Similarly, the most advanced approaches to facial feature tracking are based on Kalman (e.g., [32]) and particle filtering tracking schemes (e.g., [82]). Although face pose and facial feature tracking technologies have improved significantly in the recent years with sequential state estimation approaches that run in real time, tracking multiple, possibly occluded, expressive faces, their poses, and facial feature positions simultaneously in unconstrained environments is still a difficult problem.

The same is true for eye gaze tracking [22]. To determine the direction of the gaze, eye tracking systems employ either the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil, or computer vision techniques to find the eyes in the input image and then determine the orientation of the irises. Although there are now several companies that sell commercial eye trackers like SMI GmbH, EyeLink, Tobii, Interactive Minds, etc., realizing non-intrusive (non-wearable), fast, robust, and accurate eye tracking remains a difficult problem even in computer-centred HCI scenarios in which the user is expected to remain in front of the computer but is allowed to shift his or her position in any direction for more than 30 cm.

Because of the practical importance of the topic for affective, perceptual, and ambient interfaces of the future and theoretical interest from cognitive scientists [45], [59], automatic analysis of facial expressions attracted the interest of many researchers. Most of the facial expressions analyzers developed so far attempt to recognize a small set of prototypic emotional facial expressions such as happiness or sadness (see also the state of the art in facial affect recognition in the text below) [59]. To facilitate detection of subtle facial signals like a frown or a smile and to make facial expression information available for usage in applications like anticipatory

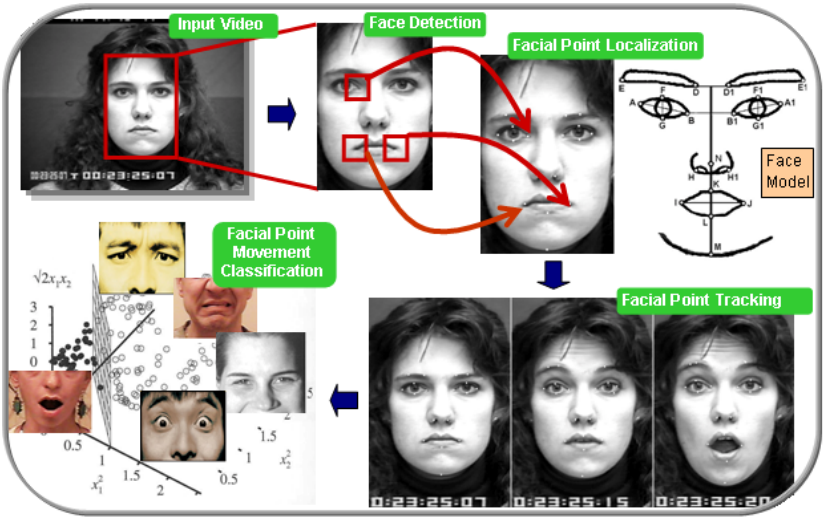


Fig. 3. Outline of an automated facial expression (AU) detection method [82]

ambient interfaces, several research groups begun research on machine analysis of facial muscle actions (atomic facial signals, action units, AUs, [25]). As AUs are independent of interpretation, they can be used for any higher order decision making process including recognition of basic emotions [25], cognitive states like interest, (dis)agreement and puzzlement [18], psychological states like suicidal depression [35] or pain [88], and social signals like emblems, regulators, and illustrators [24]. Hence, AUs are very suitable to be used as mid-level parameters in automatic facial behavior analysis, as the thousands of anatomically possible expressions can be described as combinations of 5 dozens of AUs and can be mapped to any higher order facial display interpretation [13]. A number of promising prototype systems have been proposed recently that can recognize 15 to 27 AUs (from a total of 44 AUs) in either (near-) frontal view or profile view face image sequences [79], [57]. Most of these employ statistical and ensemble learning techniques and are either feature-based (i.e., use geometric features like facial points or shapes of facial components, e.g., see Fig. 3) or appearance-based (i.e., use texture of the facial skin including wrinkles, bulges, and furrows). It has been reported that methods based on appearance features usually outperform those based on geometric features. Recent studies have shown that this claim does not always hold [58], [57]. Besides, it seems that using both geometric and appearance features might be the best choice for certain facial cues [58]. One of the main criticisms that these works received from both cognitive and computer scientists, is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify the displayed facial behavior rather than the exaggerated changes that typify posed expressions. Hence, the focus of the research in the field started to shift to automatic AU recognition in spontaneous facial expressions (produced in a reflex-like manner). Several works have recently emerged on machine analysis of AUs in spontaneous facial expression data (e.g., [14], [4], [83]). These methods use probabilistic, statistical, and ensemble learning techniques,

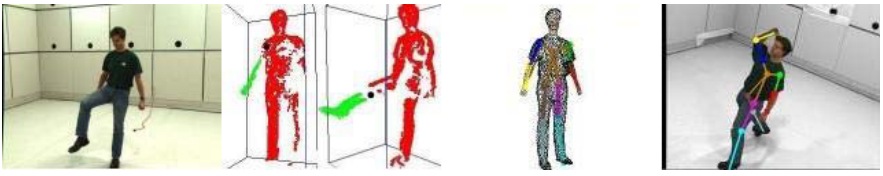


Fig. 4. Human Motion Tracking based on 3D Human Kinematic Model [11]

and perform with reasonably high accuracy in more or less constrained environments. However, the present systems for facial AU detection typically depend on relatively accurate head, face, and facial feature tracking as input and are still rather limited in performance and robustness when the input recordings are made in less constrained environments [57].

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 [85], [86], [53]. Most of the proposed techniques are either (Fig. 4, Fig. 5): model-based (i.e., use geometric primitives like cones and spheres to model head, trunk, limbs and fingers), appearance-based (i.e., use color or texture information to track the body and its parts), salient-points-based (i.e., use local signal complexity or extremes of changes in the entropy in space and time that correspond to peaks in hand or body activity variation), or spatio-temporal-shape-based (i.e., treat human body gestures as shapes in space-time domain). Most of these methods emphasize Gaussian models, probabilistic learning, and particle filtering framework (e.g., [69], [76], [53]). However, body and hands detection and tracking in unconstrained environments where large changes in illumination and cluttered or dynamic background may occur still pose significant research challenges. Also, in casual human behavior, the hands do not have to be always visible (in pockets, under the arms in a crossed arms position, on the back of the neck and under the hair), they may be in a cross fingered position, and one hand may be (partially) occluded by the other. Although some progress has been made to tackle these problems using the knowledge on human kinematics, most of the present methods cannot handle such cases correctly.

In contrast to the linguistic part of a spoken message (*what* has been said) [29], the nonlinguistic part of it (*how* it has been said) carries important information about the speaker's affective state [40] and attitude [67]. This finding instigated the research on automatic analysis of vocal nonlinguistic expressions. The vast majority of present work is aimed at discrete emotion recognition from auditory features like pitch, intensity, and speech rate (see the state of the art in vocal affect recognition in the text below) [54], [59]. For the purposes of extracting auditory features from input audio signals, freely available signal processing toolkits like Praat[1] are usually used. More recently, few efforts towards automatic recognition of nonlinguistic vocalizations like laughs [80], cries [56], and coughs [47] have been also reported. Since the research in cognitive sciences provided some promising hints that vocal outbursts and nonlinguistic vocalizations like yelling, laughing, and sobbing, may be very important

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

Fig. 5. Human body gesture recognition based on (left to right): spatio-temporal salient points, motion history (i.e., when and where motion occurred), and spatio-temporal shapes [53]

cues for decoding someone's affect/attitude [67], we suggest a much broader focus on machine recognition of these nonlinguistic vocal cues.

3.2 Context Sensing

Context plays a crucial role in understanding of human behavioral signals, since they are easily misinterpreted if the information about the situation in which the shown behavioral cues have been displayed is not taken into account [59]. For computing technology applications, context can be defined as any information that can be used to characterize the situation that is relevant to the interaction between users and the application [20]. As explained in section 2.3, six questions summarize the key aspects of the computer's context with respect to nearby humans: *Who*, *Where*, *What*, *How*, *When*, and *Why*.

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

Because of its relevance for the security, the *Who* context question has received the most attention from both funding agencies and commercial enterprises and, in turn, it has seen the most progress. The biometrics market has increased dramatically in recent years, with multiple companies providing face recognition systems like Cognitec and Identix, whose face recognition engines achieved repeatedly top 2D face recognition scores in USA government testing (FRGC, FRVT 2002, FERET 1997). The problem of face recognition has been tackled in various ways in 2D and 3D, using feature-, shape-, and appearance-based approaches as well as the combinations thereof [91], [44], [7]. The majority of the present methods employ spectral methods for dimensionality reduction like PCA, LDA, and ICA. Except of the face, biometric systems can be based on other biometric traits like fingerprints, voice, iris, retina, gait, ear, hand geometry, brainwaves, and facial thermogram [39], [64]. Biometric systems should be deployed in real-world applications and, in turn, should be able to handle a variety of problems including sensor malfunctioning, noise in sensed data, intra-class variations (e.g., facial expression which is treated as noise in face recognition), and spoof attacks (i.e., falsification attempts). Since most of these problems can be addressed by using multiple biometrics [39], [72] multimodal

Fig. 6. Multimodal biometric systems are currently a research trend [39], [72]

biometric systems have recently become a research trend (Fig. 6). The most commonly researched multi-biometrics relate to audiovisual speaker recognition. For a survey of commercial systems for alternative biometrics, see [93]. For current research efforts in multi-biometrics, see MMUA[2].

Similarly to the *Who* context question, security concerns also drive the research tackling the *Where* context-sensing problem, which is typically addressed as a computer-vision problem of surveillance and monitoring. The work in this area is based on one or more unobtrusively mounted cameras used to detect and track people. The process usually involves [86]: scene (background) modeling, motion segmentation, object classification, and object tracking. The vast majority of scene modeling approaches can be classified as generative models [9]. However, generative approaches, which require excessive amount of training data, are not appropriate for complex and incomplete problem domains like dynamic scene modeling. Unsupervised learning techniques are a better choice in that case. Motion segmentation aims at detecting regions in the scene which correspond to moving objects like cars and humans. It is one of the oldest computer vision problems and it has been tackled in various ways including [86]: background subtraction, temporal differencing, optical flow, watershed, region growing, scene mosaicing, statistical and Bayesian methods. Since natural scenes may contain multiple moving regions that may correspond to different entities, it is crucial to distinguish those that correspond to humans for the purposes of sensing the human part of the computer's context. Note that this step is superfluous where the moving objects are known to be humans. Present methods to moving object classification are usually either shape-based (e.g., human-silhouette-based) or motion-based (i.e., employ the premise that human articulated motion shows a periodic property) [86]. When it comes to human tracking for the purposes of answering the *where* context question, typically employed

[2] MMUA: http://mmua.cs.ucsb.edu/

methods emphasize probabilistic methods like Dynamic Bayesian Networks and sequential state estimation techniques like Kalman and particle filtering schemes [85], [86]. In summary, since most approaches base their analysis on segmentation and tracking, these present methods are adequate when a priori knowledge is available (e.g., the shape of the object to be tracked), but they are weak for unconstrained environments (e.g., gym, a house party), in which multiple occlusions and clutter may be present. For such cases, methods that perform analysis at the lowest semantic level (i.e., consider only temporal pixel-based behaviour) and use unsupervised learning represent a better solution (e.g., [5]).

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, head orientation, finger pointing, or simply based on the knowledge of current events like keystrokes, mouse movements, and active software (e.g., web browser, e-mail manager). However, as traditional HCI and usability-engineering applications involve relatively well-defined user tasks, many of the methods developed for user task analysis in typical HCI domains are inappropriate for task analysis in the context of human computing and ubiquitous, anticipatory ambient interfaces, where the tasks are often ill-defined due to uncertainty in the sensed environmental and behavioral cues. Analysis of tasks that human may carry out in the context of anticipatory ambient interfaces require adaptation and fusion of existing methods for behavioral cues recognition (e.g., hand/body gesture recognition, focus of attention identification) and those machine learning techniques that can be applicable to solving ill-structured decision-making problems (e.g., Markov decision processes and hidden-state models). However, only a very limited research has been directed to multimodal user's task identification in the context of anticipatory ambient interfaces and the majority of this work is aimed at support of military activities (e.g., airplane cockpit control) and crisis management [71]. Other methods for human activity recognition typically identify the task of the observed person in an implicit manner, by recognizing different tasks as different activities. The main shortcoming of these approaches is the increase of the problem dimensionality – for the same activity, different recognition classes are defined, one for each task (e.g., for the sitting activity, categories like watching TV, dining, and working with desktop computer, may be defined).

The *How* context question is usually addressed as a problem of human sensing (see the state of the art in human sensing in the text above; for a survey on speech recognition see [19]). When it comes to desktop computer application, additional modalities like writing, keystroke (choice and rate), and mouse gestures (clicks and movements) may be considered as well when determining the information that the user has passed on.

There is now a growing body of psychological research that argues that temporal dynamics of human behavior (i.e., the timing and the duration of behavioral cues) is a critical factor for interpretation of the observed behavior [67], [26]. For instance, it has been shown that facial expression temporal dynamics are essential for categorization of complex psychological states like various types of pain and mood [88] and for interpretation of social behaviors like social inhibition, embarrassment, amusement, and shame [16]. Temporal dynamics of human behavior also represent a key parameter in differentiation between posed and spontaneous human behavior. For

example, spontaneous smiles are smaller in amplitude, longer in total duration, and slower in onset and offset time than posed smiles (e.g., a polite smile) [23]. Another study showed that spontaneous smiles, in contrast to posed smiles, can have multiple apexes (multiple rises of the mouth corners – AU12) and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1s [15]. Similarly, it was shown that the differences between spontaneous and deliberately displayed brow actions (AU1, AU2, AU4) is in the duration and the speed of onset and offset of the actions and in the order and the timing of actions' occurrences [83]. In spite of these findings in basic research and except few studies on facial expression analysis (e.g., [83]), present methods for human activity/behavior recognition do not address the *When* context question: the timing of displayed behavioral signals with respect to other behavioral signals is usually not taken into account. When it comes to the timing of shown behavioral signals with respect to changes in the environment, current methods typically approach the *When* question in an implicit way, by recognizing user's reactions to different changes in the environment as different activities.

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

As can be seen from the overview of the current state of the art in so-called W5+ (Who, Where, What, When, Why, How) technology, context questions are usually addressed separately and often in an implicit manner. Yet, the context questions may be more reliably answered if they are answered in groups of two or three using the information extracted from multimodal input streams. Some experimental evidence supports this hypothesis [51]. For example, solutions for simultaneous speaker identification (*Who*) and location (*Where*) combining the information obtained by multiple microphones and surveillance cameras had an improved accuracy in comparison to single-modal and single-aspect approaches to context sensing. A promising approach to realizing multimodal multi-aspect context-sensing has been proposed by Nock et al. [51]. In this approach, the key is to automatically determine whether observed behavioral cues share a common cause (e.g., whether the mouth movements and audio signals complement to indicate an active known or unknown speaker (How, Who, Where) and whether his or her focus of attention is another person or a computer (What, Why)). The main advantages of such an approach are effective handling of uncertainties due to noise in input data streams and the problem-dimensionality reduction. Therefore, we suggest a much broader focus on spatial and temporal, multimodal multi-aspect context-sensing.

3.3 Understanding Human Behavior

Eventually, automated human behavior analyzers should terminate their execution by translating the sensed human behavioral signals and context descriptors into a

description of the shown behavior. The past work in this field can be roughly divided into the methods for understanding human affective / attitudinal states and those for understanding human social signaling (i.e., emblems, regulators, and illustrators).

Understanding Human Affect: As soon as research findings in HCI and usability engineering have suggested that HCI systems which will be capable of sensing and responding properly to human affective states are likely to be perceived as more natural, efficacious, and trustworthy, the interest in human affect machine analysis has surged. The existing body of literature in machine analysis of human affect is immense [59], [54], [57]. 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 (e.g., Fig. 7). They achieve an accuracy of 64% to 98% when detecting 3-7 emotions deliberately displayed by 5-40 subjects. However, the capabilities of these current approaches to human affect recognition are rather limited.

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

Few exceptions from this overall state of the art in the field include a few tentative efforts to detect attitudinal and non-basic affective states such as boredom [27], fatigue [32], and pain from face video [4], a few works on user-profiled interpretation of behavioral cues like facial expressions [81], and a few attempts to discern spontaneous from volitionally displayed facial behavior [83].

The importance of making a clear distinction between spontaneous and deliberately displayed human behavior for developing and testing computer systems becomes apparent when we examine the neurological substrate for facial expression. There are two distinct neural pathways that mediate facial expressions, each one originating in a different area of the brain. Volitional facial movements originate in the cortical motor strip, whereas the more involuntary, emotional facial actions, originate in the sub-cortical areas of the brain. Research documenting these differences was sufficiently reliable to become the primary diagnostic criteria for certain brain lesions prior to modern imaging methods [8]. The facial expressions mediated by these two pathways have differences both in which facial muscles are moved and in their dynamics [26]. Sub-cortically initiated facial expressions (the involuntary group) are characterized by synchronized, smooth, symmetrical, consistent, and reflex-like facial movements

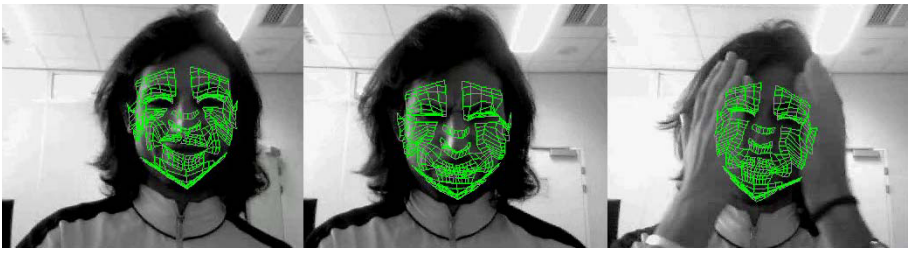


Fig. 7. 3D wireframe face-model fitting for happy, angry, and occluded face images [78]

whereas cortically initiated facial expressions are subject to volitional real-time control and tend to be less smooth, with more variable dynamics [65], [23]. Hence, having computer systems developed and tested only using deliberate (exaggerated) human behavior displays makes this technology inapplicable to real-world (naturalistic) contexts of use.

Few works in the field have been also proposed that combine several modalities into a single system for human affect analysis. Although the studies in basic research suggest that the combined face and body are the most informative for the analysis of human expressive behavior [2], only 2-3 efforts are reported on automatic human affect analysis from combined face and body gestures (e.g., [33], [42]). Existing works combining different modalities into a single system for human affective state analysis investigated mainly the effects of a combined detection of facial and vocal expressions of affective states [59], [92]. In general, these works achieve an accuracy of 72% to 85% when detecting one or more basic emotions from clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. Thus, present systems for multimodal human affect analysis have all (and some additional) limitations of single-modal analyzers. Many improvements are needed if those systems are to be used for context-sensitive analysis of naturalistic human behavior where a clean input from a known actor/ announcer cannot be expected and a context-independent processing and interpretation of audiovisual data do not suffice.

An additional important issue is that we cannot conclude that a system attaining a 92% average recognition rate performs "better" than a system achieving a 74% average recognition rate when detecting six basic emotions from audio and/or visual input stream unless both systems are tested on the same dataset. The main problem is that no audiovisual database exists that is shared by all diverse research communities in the field [59]. Although efforts have been recently reported towards development of benchmark databases that can be shared by the entire research community (e.g., [60], [33], Humaine-EU-NoE[3]), this remains an open, highly relevant issue.

Understanding Human Social Signaling: As we already remarked above, research findings in cognitive sciences tend to agree that at least some (if not the majority) of behavioral cues evolved to facilitate communication between people [38]. Types of

[3] Humaine Portal: http://emotion-research.net/wiki/Databases

messages conveyed by these behavioral cues include emblems, illustrators, and regulators, which can be further interpreted in terms of social signaling like turn taking, mirroring, empathy, antipathy, interest, engagement, agreement, disagreement, etc. Although each one of us understands the importance of social signaling in everyday life situations, and although a firm body of literature in cognitive sciences exists on the topic [2], [66], [67], and in spite of recent advances in sensing and analyzing behavioral cues like blinks, smiles, winks, thumbs up, yawns, laughter, etc. (see the state of the art in human sensing in the text above), the research efforts in machine analysis of human social signaling are few and tentative. An important part of the existing research on understanding human social signaling has been conducted at MIT Media Lab, under the supervision of Alex Pentland [62]. Their approach aims to discern social signals like activity level, stress, engagement, and mirroring by analyzing the engaged persons' tone of voice [21]. Other important works in the field include efforts towards analysis of interest, agreement and disagreement from facial and head movements [27] and towards analysis of the level of interest from tone of voice, head and hand movements [30]. Overall, present approaches to understanding social signaling are multimodal and based on probabilistic reasoning methods like Dynamic Bayesian Networks. However, most of these methods are context insensitive (key context issues are either implicitly addressed, i.e., integrated in the inference process directly, or they are ignored altogether) and incapable of handling unconstrained environments correctly. Thus, although these methods represent promising attempts toward encoding of social variables like status, interest, determination, and cooperation, which may be an invaluable asset in the development of social networks formed of humans and computers (like in the case of virtual worlds), in their current form, they are not appropriate for general anticipatory interfaces.

4 Guidelines for Future Research Efforts in the Field

According to the taxonomy of human movement, activity, and behavioral action proposed by Bobick [6], movements are low-level semantic primitives, requiring no contextual or temporal knowledge for the detection. Activities are sequences of states and movements, where the only knowledge required to recognize them relates to statistics of the temporal sequence. As can be seen from the overview of the past work done in the field, most of the work on human gesture recognition and human behavior understanding falls in this category. Human behavioral actions, or simply human behavior, are high-level semantic events, which typically include interactions with the environment and causal relationships. An important distinction between these different semantic levels of human behavior representation is the degree to which the context, different modalities, and time must be explicitly represented and manipulated, ranging from simple spatial reasoning to context-constrained reasoning about multimodal events shown in temporal intervals. However, most of the present approaches to machine analysis of human behavior are neither multimodal, nor context-sensitive, nor suitable for handling longer time scales. In our survey of the state of the field, we have tried to explicitly mention most of the existing exceptions from this rule in an attempt to motivate researchers in the field to treat the problem of

context-constrained analysis of multimodal behavioral signals shown in temporal intervals as one complex problem rather than a number of detached problems in human sensing, context sensing, and human behavior understanding. Besides this critical issue, there are a number of scientific and technical challenges that we consider essential for advancing the state of the art in the field.

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

- *Modalities:* How many and which behavioral channels like the face, the body, and the tone of the voice, should be combined for realization of robust and accurate human behavior analysis? Too much information from different channels seems to be confusing for human judges. Does this pertain in HCI?
- *Fusion:* At which abstraction level are these modalities to be fused? Humans simultaneously employ modalities of sight and sound. Does this tight coupling persists when the modalities are used for human behavior analysis, as suggested by some researchers, or not, as suggested by others? Does this depend on the machine learning techniques employed or not?
- *Fusion & Context:* While it has been shown that the *1+1>2* concept relevant to fusion of sensory neurons in humans pertain in machine context sensing 51, does the same hold for the other two concepts relevant to multimodal fusion in humans (i.e. context-dependent fusion and discordance handling)? Note that context-dependent fusion and discordance handling were never attempted.
- *Dynamics & Context:* Since the dynamics of shown behavioral cues play a crucial role in human behavior understanding, how the grammar (i.e., temporal evolvement) of human behavioral displays can be learned? Since the grammar of human behavior is context-dependent, should this be done in a user-centered manner [55] or in an activity/application-centered manner [52]?
- *Learning vs. Education:* What are the relevant parameters in shown human behavior that an anticipatory interface can use to support humans in their activities? How this should be (re-) learned for novel users and new contexts? Instead of building machine learning systems that will not solve any problem correctly unless they have been trained on similar problems, we should build systems that can be educated, that can improve their knowledge, skills, and plans through experience. Lazy and unsupervised learning can be promising for realizing this goal.

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

- *Initialization:* A large number of methods for human sensing, context sensing, and human behavior understanding require an initialization step. Since this is typically a slow, tedious, manual process, fully automated systems are the only acceptable solution when it comes to anticipatory interfaces of the future.
- *Robustness:* Most methods for human sensing, context sensing, and human behavior understanding work only in (often highly) constrained environments. Noise, fast movements, changes in illumination, etc., cause them to fail.
- *Speed:* Many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose for more sophisticated (but not always smarter) processing rather than for real time processing. A typical excuse is that according to Moore's Law we'll have faster hardware soon enough.

- *Training & Validation Issues:* United efforts of different research communities working in the field should be made to develop a comprehensive, readily accessible database of annotated, multimodal displays of human expressive behavior recorded under various environmental conditions, which could be used as a basis for benchmarks for efforts in the field. The related research questions include the following. How one can elicit spontaneous expressive behavior including genuine emotional responses and attitudinal states? How does one facilitate efficient, fast, and secure retrieval and inclusion of objects constituting this database? How could the performance of a tested automated system be included into the database? How should the relationship between the performance and the database objects used in the evaluation be defined?

5 Conclusions

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

Acknowledgements

The work of Maja Pantic and Anton Nijholt was partly supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction, publication AMIDA-4). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

1. Aarts, E.: Ambient intelligence drives open innovation. ACM Interactions, Vol. 12, No. 4 (2005) 66-68
2. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin, Vol. 111, No. 2 (1992) 256-274
3. Ba, S.O., Odobez, J.M.: A probabilistic framework for joint head tracking and pose estimation. Proc. Conf. Pattern Recognition, Vol. 4 (2004) 264-267

4. Bartlett, M.S., Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. Proc. Conf. Face & Gesture Recognition (2006) 223-230
5. Bicego, M., Cristani, M., Murino, V.: Unsupervised scene analysis: A hidden Markov model approach. Computer Vision & Image Understanding, Vol. 102, No. 1 (2006) 22-41
6. Bobick, A.F.: Movement, activity and action: The role of knowledge in the perception of motion. Philosophical Trans. Roy. Soc. London B, Vol. 352, No. 1358 (1997) 1257-1265
7. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multimodal 3D+2D face recognition. Computer Vision & Image Understanding, Vol. 101, No. 1 (2006) 1-15
8. Brodal, A.: Neurological anatomy: In relation to clinical medicine. Oxford University Press, New York, USA (1981)
9. Buxton, H. Learning and understanding dynamic scene activity: a review. Image & Vision Computing, Vol. 21, No. 1 (2003) 125-136
10. Cacioppo, J.T., Berntson, G.G., Larsen, J.T., Poehlmann, K.M., Ito, T.A.: The psychophysiology of emotion. In: Lewis, M., Haviland-Jones, J.M. (eds.): Handbook of Emotions. The Guilford Press, New York, USA (2000) 173-191
11. Cheung, K.M.G., Baker, S., Kanade, T.: Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, Vol. 1 (2003) 77-84
12. Chiang, C.C., Huang, C.J.: A robust method for detecting arbitrarily tilted human faces in color images. Pattern Recognition Letters, Vol. 26, No. 16 (2005) 2518-2536
13. Cohn, J.F.: Foundations of human computing: Facial expression and emotion. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 233-238
14. Cohn, J.F., Reed, L.I., Ambadar, Z., Xiao, J., Moriyama, T.: (2004). Automatic analysis and recognition of brow actions in spontaneous facial behavior. Proc. IEEE Int'l Conf. Systems, Man & Cybernetics (2004) 610-616
15. Cohn, J.F., Schmidt, K.L.: The timing of facial motion in posed and spontaneous smiles. J. Wavelets, Multi-resolution & Information Processing, Vol. 2, No. 2 (2004) 121-132
16. Costa, M., Dinsbach, W., Manstead, A.S.R., Bitti, P.E.R.: Social presence, embarrassment, and nonverbal behavior. Journal of Nonverbal Behavior, Vol. 25, No. 4 (2001) 225-240
17. Coulson, M.: Attributing emotion to static body postures: Recognition accuracy, confusions, & viewpoint dependence. J. Nonverbal Behavior, Vol. 28, No. 2 (2004) 117-139
18. Cunningham, D.W., Kleiner, M., Wallraven, C., Bülthoff, H.H.: The components of conversational facial expressions. Proc. ACM Int'l Symposium on Applied Perception in Graphics and Visualization (2004) 143-149
19. Deng, B.L., Huang, X.: Challenges in adopting speech recognition. Communications of the ACM, Vol. 47, No. 1 (2004) 69-75
20. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. J. Human-Computer Interaction, Vol. 16, No. 2/4 (2001) 97-166
21. Dong, W., Pentland, A.: Modeling Influence between experts. Lecture Notes on Artificial Intelligence, Spec. Vol. AI for Human Computing, Vol. 4451 (2007)
22. Duchowski, A.T.: A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments and Computing, Vol. 34, No. 4 (2002) 455-470
23. Ekman, P.: Darwin, deception, and facial expression. Annals New York Academy of sciences, Vol. 1000 (2003) 205-221
24. Ekman, P., Friesen, W.F.: The repertoire of nonverbal behavioral categories – origins, usage, and coding. Semiotica, Vol. 1 (1969) 49-98

25. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System. A Human Face, Salt Lake City, USA (2002)
26. Ekman, P., Rosenberg, E. (eds.): What the Face Reveals. Oxford University Press, Oxford, UK (2005)
27. El Kaliouby, R., Robinson, P.: Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. Proc. Int'l Conf. Computer Vision & Pattern Recognition, vol. 3 (2004) 154-
28. Fridlund, A.J.: The new ethology of human facial expression. In: Russell, J.A., Fernandez-Dols, J.M. (eds.): The psychology of facial expression. Cambridge University Press, Cambridge, UK (1997) 103-129
29. Furnas, G., Landauer, T., Gomes L., Dumais, S.: The vocabulary problem in human-system communication, Communications of the ACM, Vol. 30, No. 11 (1987) 964-972
30. Gatica-Perez, D., McCowan, I., Zhang, D., Bengio, S.: Detecting group interest level in meetings. Proc. Int'l Conf. Acoustics, Speech & Signal Processing, vol. 1 (2005) 489-492
31. Gibson, K.R., Ingold, T. (eds.): Tools, Language and Cognition in Human Evolution. Cambridge University Press, Cambridge, UK (1993)
32. Gu, H., Ji, Q.: Information extraction from image sequences of real-world facial expressions. Machine Vision and Applications, Vol. 16, No. 2 (2005) 105-115
33. Gunes, H., Piccardi, M.: Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. Proc. Int'l Conf. Systems, Man and Cybernetics (2005) 3437- 3443
34. Haykin, S., de Freitas, N. (eds.): Special Issue on Sequential State Estimation. Proceedings of the IEEE, Vol. 92, No. 3 (2004) 399-574
35. Heller, M., Haynal, V.: Depression and suicide faces. In: Ekman, P., Rosenberg, E. (eds.): What the Face Reveals. Oxford University Press, New York, USA (1997) 339-407
36. Huang, K.S., Trivedi, M.M.: Robust real-time detection, tracking, and pose estimation of faces in video. Proc. Conf. Pattern Recognition, vol. 3 (2004) 965-968
37. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. J. Computer Vision, Vol. 29, No. 1 (1998) 5-28
38. Izard, C.E.: Emotions and facial expressions: A perspective from Differential Emotions Theory. In: Russell, J.A., Fernandez-Dols, J.M. (eds.): The psychology of facial expression. Cambridge University Press, Cambridge, UK (1997) 57-77
39. Jain, A.K., Ross, A.: Multibiometric systems. Communications of the ACM, Vol. 47, No. 1 (2004) 34-40
40. Juslin, P.N., Scherer, K.R.: Vocal expression of affect. In: Harrigan, J., Rosenthal, R., Scherer, K. (eds.): The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, Oxford, UK (2005)
41. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng., Vol. 82 (1960) 35-45
42. Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., Kollias, S.: Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. Lecture Notes on Artificial Intelligence, Spec. Vol. AI for Human Computing, Vol. 4451 (2007)
43. Keltner, D., Ekman, P.: Facial expression of emotion. In: Lewis, M., Haviland-Jones, J.M. (eds.): Handbook of Emotions. The Guilford Press, New York, USA (2000) 236-249
44. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York, USA (2005)
45. Lisetti, C.L., Schiano, D.J.: Automatic facial expression interpretation: Where human-computer interaction, AI and cognitive science intersect. Pragmatics and Cognition, Vol. 8, No. 1 (2000) 185-235

46. Maat, L., Pantic, M.: Gaze-X: Adaptive affective multimodal interface for single-user office scenarios. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 171-178
47. Matos, S., Birring, S.S., Pavord, I.D., Evans, D.H.: Detection of cough signals in continuous audio recordings using HMM. IEEE Trans. Biomedical Engineering, Vol. 53, No. 6 (2006) 1078-1083
48. Nijholt, A., Rist, T., Tuinenbreijer, K.: Lost in ambient intelligence. Proc. Int'l Conf. Computer Human Interaction (2004) 1725-1726
49. Nijholt, A., de Ruyter, B., Heylen, D., Privender, S.: Social Interfaces for Ambient Intelligence Environments. In: Aarts, E., Encarnaçao, J. (eds.): True Visions: The Emergence of Ambient Intelligence. Springer, New York, USA (2006) 275-289
50. Nijholt, A., Traum, D.: The Virtuality Continuum Revisited. Proc. Int'l Conf. Computer Human Interaction (2005) 2132-2133
51. Nock, H.J., Iyengar, G., Neti, C.: Multimodal processing by finding common cause. Communications of the ACM, Vol. 47, No. 1 (2004) 51-56
52. Norman, D.A.: Human-centered design considered harmful. ACM Interactions, Vol. 12, No. 4 (2005) 14-19
53. Oikonomopoulos, A., Patras, I., Pantic, M., Paragios, N.: Trajectory-based Representation of Human Actions. Lecture Notes on Artificial Intelligence, Spec. Vol. AI for Human Computing, Vol. 4451 (2007)
54. Oudeyer, P.Y.: The production and recognition of emotions in speech: features and algorithms. Int'l J. Human-Computer Studies, Vol. 59, No. 1-2 (2003) 157-183
55. Oviatt, S.: User-centered modeling and evaluation of multimodal interfaces. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1457-1468
56. Pal, P., Iyer, A.N., Yantorno, R.E.: Emotion detection from infant facial expressions and cries. Proc. Int'l Conf. Acoustics, Speech & Signal Processing, Vol. 2 (2006) 721-724
57. Pantic, M., Bartlett, M.S.: Machine Analysis of Facial Expressions. In: K. Kurihara (ed.): Face Recognition. Advanced Robotics Systems, Vienna, Austria (2007)
58. Pantic, M., Patras, I.: Dynamics of Facial Expressions – Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. IEEE Trans. Systems, Man, and Cybernetics, Part B, Vol. 36, No. 2 (2006) 433-449
59. Pantic, M., Rothkrantz, L.J.M.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1370-1390
60. Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. Proc. IEEE Int'l Conf. Multimedia and Expo (2005) 317-321 (www.mmifacedb.com)
61. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. Proc. IEEE Int'l Conf. Face and Gesture Recognition (2004) 97-102
62. Pentland, A. Socially aware computation and communication. IEEE Computer, Vol. 38, No. 3 (2005) 33-40
63. Pitt, M.K., Shephard, N.: Filtering via simulation: auxiliary particle filtering. J. Amer. Stat. Assoc., Vol. 94 (1999) 590-599
64. Prabhakar, S., Kittler, J., Maltoni, D., O'Gorman, L., Tan, T.: Introduction to the Special Issue on Biometrics: Progress and Directions. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, No. 4 (2007) 513-516
65. Rinn, W. E.: The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. Psychological Bulletin, Vol. 95, No. 1 (1984) 52–77
66. Russell, J.A., Fernandez-Dols, J.M. (eds.): The psychology of facial expression. Cambridge University Press, Cambridge, UK (1997)

67. Russell, J.A., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and Vocal Expressions of Emotion. Annual Review of Psychology, Vol. 54 (2003) 329-349
68. Ruttkay, Z.M., Reidsma, D., Nijholt, A.: Human computing, virtual humans, and artificial imperfection. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 179-184
69. Sand, P., Teller, S.: Particle Video: Long-Range Motion Estimation using Point Trajectories. Proc. Int'l Conf. Computer Vision and Pattern Recognition (2006) 2195-2202
70. Scanlon, P., Reilly, R.B.: Feature analysis for automatic speech reading. Proc. Int'l Workshop Multimedia Signal Processing (2001) 625-630
71. Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Maceachren, A.M., Sengupta, K.: Speech-gesture driven multimodal interfaces for crisis management. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1327-1354
72. Sim, T., Zhang, S., Janakiraman, R., Kumar, S.: Continuous Verification Using Multimodal Biometrics. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, No. 4 (2007) 687-700
73. Song, M., Bu, J., Chen, C., Li, N.: Audio-visual based emotion recognition – A new approach. Proc. Int'l Conf. Computer Vision and Pattern Recognition (2004) 1020-1025
74. Starner, T.: The Challenges of Wearable Computing. IEEE Micro, Vol. 21, No. 4 (2001) 44-67
75. Stein, B., Meredith, M.A.: The Merging of Senses. MIT Press, Cambridge, USA (1993)
76. Stenger, B., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical Bayesian filter. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 28, No. 9 (2006) 1372-1384
77. Streitz, N., Nixon, P.: The Disappearing Computer. ACM Communications, Vol. 48, No. 3 (2005) 33-35
78. Tao, H., Huang, T.S.: Connected vibrations – a model analysis approach to non-rigid motion tracking. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (1998) 735-740
79. Tian, Y.L., Kanade, T., Cohn, J.F.: Facial Expression Analysis. In: Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York, USA (2005) 247-276
80. Truong, K.P., van Leeuwen, D.A.: Automatic detection of laughter. Proc. Interspeech Euro. Conf. (2005) 485-488
81. Valstar, M.F., Pantic, M.: Biologically vs. logic inspired encoding of facial actions and emotions in video. Proc. IEEE Int'l Conf. on Multimedia and Expo (2006) 325-328.
82. Valstar, M.F., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, vol. 3 (2006) 149-
83. Valstar, M.F., Pantic, M., Ambdar, Z., Cohn, J.F.: Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 162-170
84. Viola, P., Jones, M.J.: Robust real-time face detection. Int'l J. Computer Vision, Vol. 57, No. 2 (2004) 137-154
85. Wang, J.J., Singh, S.: Video analysis of human dynamics – a survey. Real Time Imaging, Vol. 9, No. 5 (2003) 321-346
86. Wang, L., Hu, W., Tan, T.: Recent developments in human motion analysis. Pattern Recognition, Vol. 36, No. 3 (2003) 585-601
87. Weiser, M.: The Computer for the Twenty-First Century. Scientific American, Vol. 265, No. 3 (1991) 94-104
88. Williams, A.C.: Facial expression of pain: An evolutionary account. Behavioral & Brain Sciences, Vol. 25, No. 4 (2002) 439-488

89. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 1 (2002) 34-58
90. Zhai, S., Bellotti, V.: Sensing-Based Interaction. ACM Trans. Computer-Human Interaction, Vol. 12, No. 1 (2005) 1-2
91. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys, Vol. 35, No. 4 (2003) 399-458
92. Zeng, Z., Hu, Y., Roisman, G.I., Fu, Y., Huang, T.S.: Audio-visual Emotion Recognition in Adult Attachment Interview. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 139-145
93. BTT Survey on Alternative Biometrics. Biometric Technology Today, Vol. 14, No. 3 (2006) 9-11

Audio-Visual Spontaneous Emotion Recognition

Zhihong Zeng[1], Yuxiao Hu[1], Glenn I. Roisman[1], Zhen Wen[2], Yun Fu[1], and Thomas S. Huang[1]

[1] University of Illinois at Urbana-Champaign, USA
[2] IBM T.J.Watson Research Center, USA
{zhzeng, hu3, yunfu2, huang}@ifp.uiuc.edu, roisman@uiuc.edu, zhenwen@us.ibm.com

Abstract. Automatic multimodal recognition of spontaneous emotional expressions is a largely unexplored and challenging problem. In this paper, we explore audio-visual emotion recognition in a realistic human conversation setting—the Adult Attachment Interview (AAI). Based on the assumption that facial expression and vocal expression are at the same coarse affective states, positive and negative emotion sequences are labeled according to Facial Action Coding System. Facial texture in visual channel and prosody in audio channel are integrated in the framework of Adaboost multi-stream hidden Markov model (AdaMHMM) in which the Adaboost learning scheme is used to build component HMM fusion. Our approach is evaluated in AAI spontaneous emotion recognition experiments.

Keywords: Multimodal Human-Computer Interaction, Affective computing, affect recognition, emotion recognition.

1 Introduction

Human-computer interaction has been a predominantly one-way interaction where a user needs to directly request computer responses. Change in the user's affective state, which play a significant role in perception and decision making during human to human interactions, is inaccessible to computing systems. Emerging technological advances are enabling and inspiring the research field of "affective computing," which aims at allowing computers to express and recognize affect [1]. The ability to detect and track a user's affective state has the potential to allow a computing system to initiate communication with a user based on the perceived needs of the user within the context of the user's actions. In this way, human computer interaction can become more natural, persuasive, and friendly [2-3][45][65].

In the speech recognition community, there is an increasing interest in improving performance of spontaneous speech recognizers by taking into account the influence of emotion on speech [5-7]. The authors in [7] made a systematic comparison of speech recognition under different conditions. Their results show that the influence of emotion is larger than others (i.e. noise, loudness). The studies [5-6] indicated that emotion-sensitive audio-only ASR system improved speech recognition rates noticeably.

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 72–90, 2007.

Automatic emotion recognition has been attracting attention of researchers from a variety of different disciplines. Another application of automatic emotion recognition is to help people in emotion-related research to improve the processing of emotion data. In recent decades, with the advance of emotion theories [10][11][12] and emotion measurement (e.g. Facial Action Unit System (FACS) [13]), more and more reports of emotion analysis have been conducted in psychology, psychiatry, education, anthropology, neurophysiology [14][15][16]. The emotion-related research includes attachment [17], mother-infant interaction [18], tutoring [19], and psychiatric disorders [20]. All of the above research requires measurement of emotion expressions. At present, this problem was solved mainly by self-reports and observers' judgments of emotion (based on FACS or other labeling schemes). But self-reports and human-based emotion measurements are error-prone, and time consuming. Such limitations influence the rate at which new research can be done. Automatic emotion recognition would reduce dramatically the time it takes to measure emotional states, and improve the reliability of measurement.

In this paper, we explore audio-visual recognition of spontaneous emotions occurring in a realistic human conversation setting—the Adult Attachment Interview (AAI). The AAI is the most widely used and well-validated instrument in developmental research for identifying adult attachment representations. The AAI data in our experiment were collected by the authors in [17] to study links between adults' narratives about their childhood experiences and their emotional expressive, physiological, and self-reported emotion.

Although the ability to recognize a variety of fine-grained emotions is attractive, it may be not practical because the emotional data in the context of realistic conversations is often not sufficient to learning a classifier for a variety of fine-grained emotions. In this paper, we focus on recognizing positive and negative emotions which can be used as a strategy to improve the quality of interface in HCI, and as a measurement in studies conducted in the field of psychology [17]. This work extends our previous work [21][55] that explored separating spontaneous emotional facial expressions from non-emotional facial expressions in order to narrow down data of interest for emotion recognition research.

In the paper, we propose Adaboost multi-stream hidden Markov model (Adaboost MHMM) to integrate audio and visual affective information. In order to capture the richness of facial expression, we use 3D face tracker to extract facial texture images that are then transformed into low dimensional subspace by Locality Preserving Projection (LPP). We use pitch and energy in audio channel to build audio HMM in which some prosody features, like frequency and duration of silence, could have implications. In the audiovisual fusion stage, we treat the component HMM combination as a multi-class classification problem in which the input is the probabilities of HMM components and the output is the target classes, based on the training combination strategy [44]. We use Adaboost learning scheme to build fusion of the component HMMs from audio and visual channels. The framework of our approach is illustrated in Figure 1.

The rest of the paper is organized as follows. In the following section, we briefly describe related work about automatic emotion recognition, focusing on audio-visual emotion recognition. In Section 3 we introduce the AAI data that is used in our spontaneous affective expression analysis. Section 4 introduces a 3D face tracker used

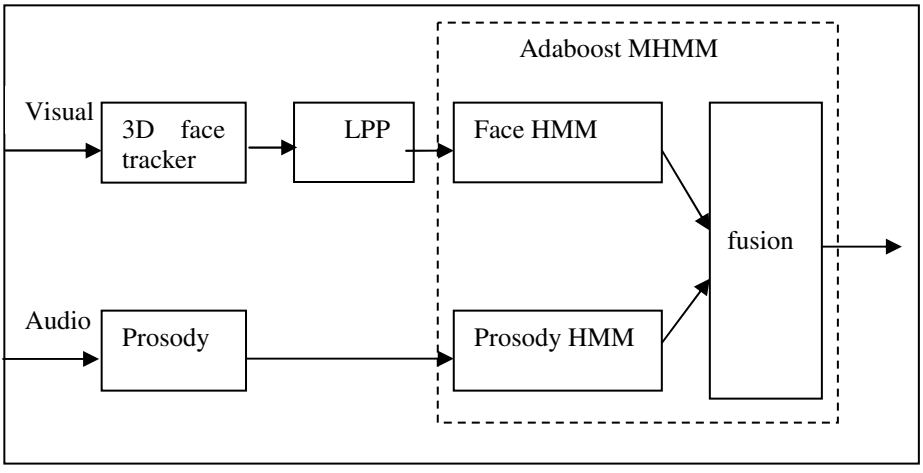


Fig. 1. The audio-visual emotion recognition framework

to extract facial textures, and Locality Preserving Projection for feature dimension reduction. Section 5 describes prosodic feature extraction in audio channel. In Section 6, we introduce Adaboost multi-stream hidden Markov model for bimodal fusion. Section 7 presents our preliminary experimental results on two AAI subjects to evaluate our method. Finally, we have concluding remarks in Section 8.

2 Related Work

The recent appraisal-based emotion theory [12] indicates the importance of the integration of information from different response components (such as facial and vocal expression) to yield a coherent judgment of emotions. In addition, current techniques of both computer vision and audio processing have limitations in realistic applications. For instance, current face trackers are sensitive to head pose, occlusion and lighting changes while audio processing is sensitive to noise and distance between speakers and microphone. Audio and visual channels provide complementary information. Moreover, if one channel fails for some reason, the other channel can still work. Thus, the final fusion performance can be robust.

In recent years, the literature on automatic emotion recognition has been growing dramatically due to the development of techniques in computer vision, speech analysis, and machine learning. However, automatic recognition on emotions occurring in natural communication settings is a largely unexplored and challenging problem. Authentic emotional expressions are difficult to collect because they are relatively rare and short lived, and filled with subtle context-based changes that make it difficult to elicit emotions without influencing the results. Manual labeling of spontaneous emotional expressions for ground truth is very time consuming, error prone, and expensive [23]. This state of affairs leads to a big challenge for spontaneous emotional expression analysis. Due to these difficulties in emotional expression recognition, most of current automatic facial expression studies have been

based on the artificial material of deliberately expressed emotions that were collected by asking the subjects to perform a series of emotional expressions to a camera. The popular artificial facial expression databases are Ekman-Hager data [24], and Kanade-Cohn data [25], Pantic et al.'s data [26], and the JAFFE data [27]. The audio-visual emotion data is Chen-Huang data [28]. An overview of databases for emotion classification studies can be founded in [26][29].The psychological study [30] indicates that the posed nature of the emotions may differ in appearance and timing from corresponding performances in natural settings. More specifically, the study [31] indicated that posed smiles were of larger amplitude and has less consistent relation between amplitude and duration than spontaneous smile. And the study [32] indicated the spontaneous brow actions (AU1, AU2 and AU4 in the Facial Action Coding System) have different characteristics (intensity, during and occurrence order) from corresponding posed brow actions. In addition, most of current emotion recognizers are evaluated in clear and constrained input (e.g., high quality visual and audio recording, non-occluded and front-view face), which is different from the natural communication setting. Therefore, the methods based these artificial emotions would be expected to perform inadequately on emotional expressions occurring in natural communication settings.

Most studies of automatic emotion recognition focus on six basic facial expressions or a subset of them, namely happiness, sadness, anger, fear, surprise, and disgust. The recognition of these basic facial expressions was based on Ekman's extensive study [33] that indicated the universality of human perception of these basic facial expressions in different cultures. However, most of the emotion expressions that occur in our human-human conversation are non-basic emotions [15].

In addition, most of current automatic emotion recognition approaches are uni-modal: information processed by the computer system is limited to either face images [66-72] or speech signals [73-77]. Relatively little work has been done in researching multimodal affect analysis. For extensive survey of automatic emotion analysis done in the recent years, readers are referred to review papers, including [34][35][78] written by Pantic et al. in 2003, 2005 and 2006, [37] by Cowie et al. in 2001, and [36] by Sebe et al. in 2005.

Here we focus on reviewing the efforts toward audio-visual emotion recognition, especially those done in these years, which have not been included in the previous review papers. The first studies for audio-visual emotion recognition include Chen and Huang [38][39], Yoshitomi et al. [40], De Silva and Ng [41]. In that past few years, there are an increasing number of reports investigating the integration of emotion-related information from audio and visual channels in order to improve the recognition performance. They are summarized in Table 1 in which the first column is the reference, the second column is the number of subjects in datasets, the third column is the number of emotional states, the fourth column is classifier, and the fifth column is fusion method (feature-level, model-level, and decision-level), the sixth column is the test method (p: person-dependent; i: person-independent), and last column is recognition accuracy (%). * denotes the missing entry. The data in [44-46][48] included 6 basic emotions (happiness, sadness, fear, disgust, anger, and surprise), 4 cognitive/ motivational states (interest, boredom, puzzlement and frustration), and neutral. The data in [51] were collected in a standing car with webcam and an array microphone. In the study [52], the data include six different languages (English, Chinese, Urdu,

Table 1. Properties of studies of Audio-visual emotion recognition on posed emotions

reference	subject	state	classifier	fusion	test	accuracy
Zeng et al. [44]	20	11	MFHMM	model	i	83.64%
Zeng et al. [45]	20	11	MFHMM	model	i	80.61%
Zeng et al. [46]	20	11	MHMM	decision	i	75%
Zeng et al. [47]	20	2	Fisher-boosting	feature	p	>84%
Zeng et al. [48]	20	11	SNoW MHMM	decision	P i	96.30% 72.42%
Song et al. [49]	*	7	THMM	model	*	84.7%
Buss et al. [50]	1	4	SVC	Feature, decision	P	89.1% 89.0%
Hoch et al. [51]	7	3	SVM	decision	p	90.7%
Wang et al. [52]	8	6	LDA	decision	i	82.14%
Go et al. [53]	10	6	LDA K-mean	decision	i	>95%

Punjabi, Persian, and Italian) and subjects were from different races. All of those studies were evaluated on artificial material of deliberately expressed emotions.

Because more and more researchers have noticed the potential difference between posed emotion expression and spontaneous emotion expression, there is in the last few years a trend in the emotion recognition community in moving away from posed emotion recognition to spontaneous emotion recognition. These notable studies include visual spontaneous expression recognition, audio spontaneous emotion recognition, and relatively few audio-visual spontaneous emotion recognition. They are summarized in Table 2 in which the first column is the reference, the second column "sub" is the number of subjects in their dataset, the third column "state" is the number of emotional states or facial action units, the fourth column "labeling" is the labeling methods, the fifth column "fea" is the used feature (A: audio; V: visual), the sixth column is the classifier, the seventh column is test method (p: person-dependent; i: person-independent), the last column is the recognition accuracy.

The datasets of these studies of spontaneous emotion expressions include human-human interaction and human-to-computer interaction. They are collected from call center [60][61], meeting [60], interview [54][56][57][32], dialogue system [56], Wizard of OZ scenarios [62][63][59], and kiosk [55].

There are methodological differences between the studies of posed emotion recognition and those of spontaneous emotion recognition. First, as compared to posed emotion recognition where labels are pre-defined, the spontaneous emotion labeling for ground truth is error-prone and time consuming. In Table 2, only the study [55] used the self-reports of subjects as labels, and the other studies used the human observation in which the humans labeled the data based on Facial Action Unit System (FACS), Feeltrace system [64], or ad hoc labeling scheme.

Second, in these studies of spontaneous emotion recognition, only one study [55] tried to recognize the 4 basic emotions (neutral, happiness, surprise and disgust). In the rest studies, the recognized emotion states include coarse emotion states (positive,

Table 2. Properties of studies of Audio-visual spontaneous emotion recognition

reference	sub	state	labeling	fea	classifier	test	accuracy
Barlett et al. [54]	12	16AUs	FACS	V	SVM	i	90.5%
Sebe et al. [55]	28	4	Self-report	V	KNN	*	95.57%
Zeng et al. [56]	2	2	FACS	V	KSVDD	p	79%
Cohn et al. [57]	21	3AUs	FACS	V	Guassian	i	76%
Valstar et al. [32]	*	3AUs	FACS	V	SVM	i	50.4%
Litman et al. [58]	10	3	ad hoc	A	decision tree	p i	66-73%
Batliner et al. [59]	24	2	ad hoc	A	LDA	i	74.2%
Neiberg et al.[60]	*	3	ad hoc	A	GMM	i	85-95%
Ang et al. [61]	*	6	ad hoc	A	decision tree	i	85.4-93.2%(2class)
Fragopanagos et al. [62]	*	4	Feel-trace	AV	Neural Network	*	44-71%
Garidakis et al. [63]	*	4	Feel-trace	AV	Neural network	*	79%

negative and neutral) [56][60], quadrant states in evaluation-activation space [62][63], or application-dependent states (trouble in [59], annoyance and frustration [61]), and facial action units [54][57][32]. The main reason could be that there is no sufficient data of basic emotion expressions to train a classifier.

Because these above studies evaluate their algorithms on the different experimental condition (data, labeling, the number of classes, feature set), it is difficult to give their performance rank only based on the accuracies.

In the studies [62][63] toward audio-visual spontaneous emotion recognition, the authors uses Feeltrace tool [64] to label the data collected in a "Wizard of OZ" scenario. In the study [62], due to considerable variation across four raters, it is difficult to reach a similar assessment with the FeelTrace labels. They observed the difference of the labeling results among four Feeltrace users. Specifically, these Feeltracers judged the emotional states of data by using different modalities. For example, one used facial expressions as the most important cues to make the decision while another used prosody.

Compared with Feeltrace tool mentioned above, FACS could be more objective in labeling and be able to capture the richness and complexity of emotional expressions. Thus, we build our emotion recognizer with FACS labeling in this work. We make the assumption that in our database there is no blended emotions so that the facial expression and prosody belong to same emotional states at the coarse level (i.e. positive and negative emotions).

In addition, different from these two studies above mentioned [62][63], we apply 3D face tracker which is able to capture the wider range of face movement than 2D face tracker. We use facial texture instead of sparse geometrical features in order to capture the richness of subtle facial expressions. For capturing the dynamic structure of emotional expressions and integrating audio and visual streams, we build our recognizer in Adaboost multi-stream hidden Markov model framework in which Adaboost learning scheme is used to build fusion of component HMMs.

3 Data of Adult Attachment Interview

The Adult Attachment Interview (AAI) is a semi-structured interview used to characterize individuals' current state of mind with respect to past parent-child experiences. This protocol requires participants to describe their early relationships with their parents, revisit salient separation episodes, explore instances of perceived childhood rejection, recall encounters with loss, describe aspects of their current relationship with their parents, and discuss salient changes that may have occurred from childhood to maturity [17].

During data collection, remotely controlled, high-resolution (720*480) color video cameras recorded the participants' and interviewer's facial behavior during AAI. Cameras were hidden from participants' view behind a darkened glass on a bookshelf in order not to distract the participant's attention. The snapshot of an AAI video is shown in Figure 2. The participant's face is displayed in the bigger window while the interviewer's face is in the smaller left-top window.

As our first step to explore audio-visual spontaneous emotion recognition, AAI data of two subjects (one female and one male) was used in this study. The video of the female subject lasted 39 minutes, and one of the male lasted 42 minutes. The significant amount of data allowed us personal-dependent spontaneous emotion analysis.

In order to objectively capture the richness and complexity of facial expressions, Facial Action Coding System (FACS) was used to code every facial event that occurred during AAI by two certified coders. Inter-rater reliability was estimated by the ratio of the number of agreements in emotional expression to the total number of agreement and disagreements, yielding for this study a mean agreement ratio of 0.85.

To reduce FACS data further for analysis, we manually grouped combinations of AUs into two coarse emotion categories (i.e., positive and negative emotions) on the basis of an empirically and theoretically derived Facial Action Coding System Emotion Codes which was created by the psychological study [79].

In order to narrow down the inventory to potential useful emotion data for our experiment, we first ignore the emotion occurrences to which these two coders disagree with each other. In order to analyze the emotions occurring in a natural communication setting, we have to face the technique challenges to handle arbitrary head movement. Due to the technique limitation, we filtered out the emotion segments in which hand occluded the face, face turned away more than 40 degree with respect to the optical center, or part of face moved out of camera view. Each emotion sequence starts from and to the emotion intensity scoring scale B (slight) or C (marked pronounced) defined in [13]. The number of audio-visual emotion expression segments in which subjects displayed emotions using both facial expressions and voice is 67 for female and 70 for male.

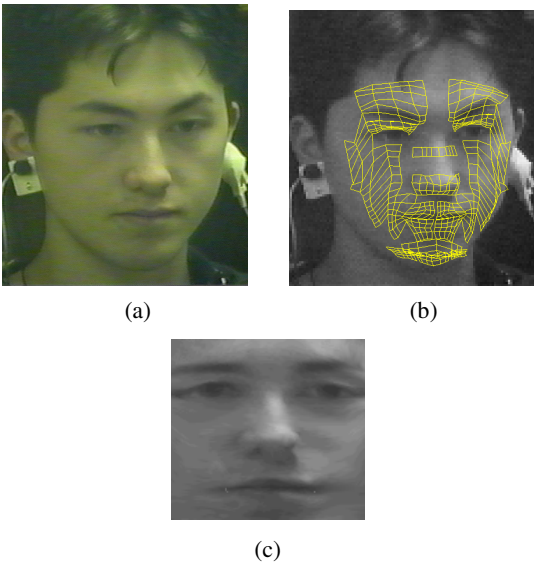
Fig. 2. The snapshot of an AAI video. The participant's face is displayed in the bigger window while the interviewer's face is in the smaller left-top window.

4 Facial Expressions

This section includes 3D face tracker and Locality Preserving Projection which aims to project the high-dimensional images to low dimensional subspace.

4.1 3D Face Tracker

To handle the arbitrary behavior of subjects in the natural setting, it is required to track the 3D face. The face tracking in our experiments is based on a system called Piecewise Bezier Volume Deformation (PBVD) tracker which was developed in [80][81].



(a) (b)

(c)

Fig. 3. The 3D face tracker's result. (a) the video frame input; (b) tracking result where a mesh is used to visualize the geometric motions of the face; (c) extracted face texture.

This face tracker uses a 3D facial mesh model which is embedded in multiple Bezier volumes. The shape of the mesh can be changed with the movement of the control points in the Bezier volumes, which guarantees the surface patches to be continuous and smooth. In the first video frame, the 3-D facial mesh model is constructed by selection of landmark facial feature points. Once the model was fitted, the tracker can track head motion and local deformations of the facial features by an optical flow method. In this study, we use rigid setting of this tracker to extract facial expression texture. The 3D ridge geometric parameters (3D rotation and 3D translation) determine the registration of each image frame to the face texture map, which is obtained by wrapping the 3D face appearance. Thus, we can derive a sequence of face texture images, which capture the richness and complexity of facial expression. Figure 3 shows a snapshot of the tracking system. Figure 3(a) is the input video frame, and Figure 3(b) is the tracking result where a mesh is used to visualize the geometric motions of the face. The extracted face texture is shown in Figure 3(c).

4.2 Locality Preserving Projection

In recent years, computer vision research has witnessed a growing interest in subspace analysis techniques. Before we utilize any classification technique, it is beneficial to first perform dimensionality reduction to project an image into a low dimensional feature space, due to the consideration of learnability and computational efficiency.

Locality Preserving Projection (LLP) is a linear mapping that is obtained by finding the optimal linear approximations to the eigen-functions of the Laplace Beltrami operator on the manifold [82]. As contrasted with nonlinear manifold learning techniques, LLP can be simply applied to any new data point to locate it in the reduced representation manifold subspace, which is suitable for classification application.

Some traditional subspace methods such as PCA aim to preserve the global structure. However, in many real world applications, especially facial expression recognition, the local structure could be more important. In contrast to PCA, LPP finds an embedding that preserves local information, and obtains a subspace that best detects the essential manifold structure.

The details of LPP, including its learning and mapping algorithms, can be found in [82]. The low-dimensional features from LPP are then used to build visual HMM.

5 Vocal Expressions

In our work, we use prosodic features which are related with the way the sentences are spoken. For audio feature extraction, Entropic Signal Processing System named get_f0, a commercial software package, is used. It implements a fundamental frequency estimation algorithm using the normalized cross correlation function and dynamic programming. The program can output the pitch F0 for fundamental frequency estimate, RMS energy for local root mean squared measurements, prob_voice for probability of voicing, and the peak normalized cross-correlation value that was used to determine the output F0. The experimental results in our previous work [48] showed pitch and energy are the most important factors in affect

classification. Therefore, in our experiment, we only used these two audio features for affect recognition. Some prosody features, like frequency and duration of silence, could have implication in the HMM structure of energy and pitch.

6 Adaboost Multi-stream Hidden Markov Model

Audio-visual fusion is an instance of the general classifier fusion problem, which is an active area of research with many applications, such as Audio-Visual Automatic Speech Recognition (AVASR). Although there are some audio-visual fusion studies in audio-visual Automatic Speech Recognition literature [24], few studies are found for audio-visual affect recognition shown in Table 1.

Most of current multi-stream combination studies focus on weighting combination scheme with weights proportional to the reliabilities of the component HMMs. The weights can be computed from normalized stream recognition rate [22], stream S/N ratio [22], stream entropy [8], or other reliability measures such as ones in [9].

The weighting combination scheme is intuitive and reasonable in some ways. However, it is based on the assumption that the combination is linear. This assumption could be invalid in practice. In addition, using the weighting scheme is difficult to obtain the optimal combination because they deal with different feature spaces and different models. It is even possible that the weighting combination is worse than individual component performance, as shown in our experiments.

According to training combination strategy in our previous work [44], the component HMM combination can be treated as a multi-class classification problem in which the input is the probabilities of HMM components and the output is the target classes. This combination mechanism can be linear or nonlinear, depending on learning scheme that we use. In this case, if s represents the number of possible classes and n the number of streams, this classification contains s×n input units and s output units, and the parameters of the classifier can be estimated by training. Under this strategy, we propose Adaboost MHMM in which the Adaboost learning scheme is used to build the fusion of multiple component HMMs.

6.1 Learning Algorithm

Given m training sequences each of which has n streams

$$(x_{11},\cdots,x_{1n},y_1),\cdots,(x_{m1},\cdots,x_{mn},y_m)$$

where x_{ij} is the jth stream of ith sample sequence, and $y_i = 0,1$ for negative and positive emotions in our application. Assume that these n streams can be modeled respectively by n component HMMs. The learning algorithm of the Adaboost MHMM includes three main steps.

1. n component HMMs are trained independently by the EM algorithm. The model parameters (the initial, transition, and observation probabilities) of individual HMMs are estimated.

2. For each training sequence, likelihoods of these n component HMMs are computed. We obtain

$$(p_{110}, p_{111}, \cdots, p_{1n0}, p_{1n1}, y_1), \cdots, (p_{m10}, p_{m11}, \cdots, p_{mn0}, p_{mn1}, y_m)$$

where p_{ij0}, p_{ij1} are likelihoods of negative and positive emotions of jth stream of ith sample sequence.
3. Fusion training: based on Adaboost learning scheme [4], these estimated likelihoods of n component HMMs are used to construct a strong classifier which is a weighted linear combination of a set of weak classifiers.

6.2 Classification Algorithm

Given a n-stream observation sequence and the model parameters of Adaboost MHMM, the inference algorithm of the Adaboost MHMM includes two main steps.

1. Compute individually likelihoods of positive and negative emotions of n component HMMs.
2. A set of weaker hypotheses are estimated each using likelihood of positive or negative emotion of a single component HMM. The final hypothesis is obtained by weighted linear combination of these hypotheses where the weights are inversely proportional to the corresponding training errors [4].

7 Experimental Results

In this section, we present the experimental results of our emotion recognition by using audio and visual affective information.

The personal-dependent recognition is evaluated on the two subjects (one female and one male) in which the training sequences and test sequences were taken from the same subject. For this test, we apply leave-one-sequence-out cross-validation. For each subject in this test, one sequence among all of emotion expression sequences is used as the test sequence, and the remaining sequences are used as training sequences. This test is repeated, each time leaving a different sequence out.

7.1 Facial Expression Analysis on Locality Preserving Subspace

In this experiment, we evaluate the LPP HMM method which models the facial expressions by using the low-dimensional features in the locality preserving subspace of facial texture images. In addition, the PCA HMM method, which uses the features in the PCA subspace, is also tested to make the performance comparison with LPP HMM. The comparison results are shown in Table 3 for these two subjects. The corresponding facial expression subspaces are called optimal facial expression subspaces for each method. Figure 4 and 5 shows a plot of recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction for female and male respectively. It is shown that LPP HMM method largely outperforms PCA HMM. The recognition accuracy of LPP HMM is 87.50% at 5D subspace for female and 84.85% at 4D subspace for male respectively.

Table 3. Performance Comparison of LPP HMM and PCA HMM

	Approach	Dimension	Accuracy (%)
Female	LPP HMM	5	87.50
	PCA HMM	5	79.69
Male	LPP HMM	4	84.85
	PCA HMM	5	72.62

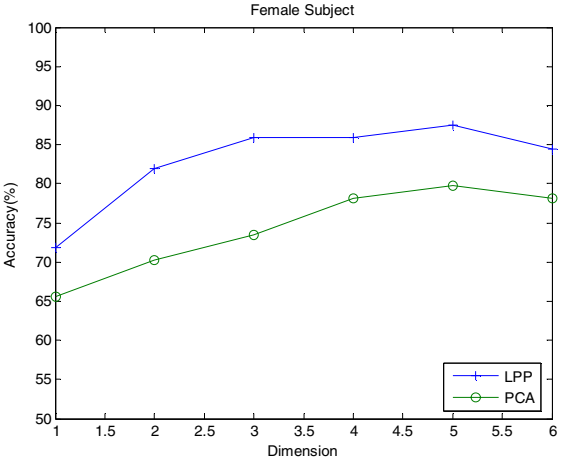


Fig. 4. Facial expression recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction on the female emotion data

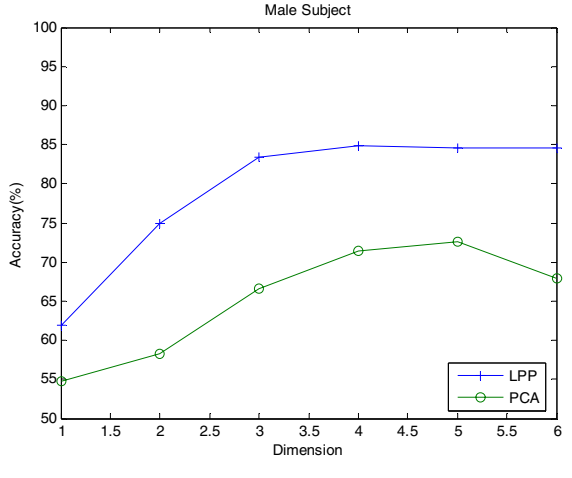


Fig. 5. Facial expression recognition accuracy of LPP HMM and PCA HMM vs. dimensionality reduction on the male emotion data

7.2 Prosody Expression Analysis

Table 4 shows the experimental results of emotion recognition by using audio HMM. The recognition performance of prosody HMM is better than random, but worse than facial expression recognition. There are two possible reasons why prosodic affective recognition is worse than facial affective recognition. One is that facial expression could provide more reliable affective information than prosody, as the psychological study indicated [15]. The other reason is that we only use information of facial expressions to label our multimedia emotion data. Thus, facial expression recognition is more agreement with human judgment (labels) than prosody expressions.

Table 4. Emotion Recognition of Prosody HMM

Subjects	Accuracy (%)
Female	75.09
Male	65.15

7.3 Audio-Visual Fusion

The emotion recognition performance of audio-visual fusion is shown in Table 3. In this table, two combination schemes (weighting and training) are used to fuse the component HMMs from audio and visual channels. Acc MHMM reflects MHMM with the weighting combination scheme in which the weights are proportional to stream normalized recognition accuracies. Adaboost MHMM mean MHMM with the Adaboost learning schemes as described in Section 6. Because we treat the multi-stream fusion as a multi-class classification problem, there are a variety of methods that can be used to build the fusion. In addition to Adaboost MHMM, we used LDC and KNN (K=3 for female and K=5 for male) to build this audio-visual fusion, which are Ldc MHMM and Knn MHMM in Table 3.

The performance comparison of these fusion methods is as follows:

Adaboost MHMM > Knn MHMM > Acc MHMM > Ldc MHMM

The results demonstrate that training combination outperforms weighting combination, except Ldc MHMM that is a linear fusion. Adaboost MHMM is the best among these four fusion methods.

The summarization of the performance of different modalities and different fusion methods is illustrated in Figure 4. Results show that Adaboost MHMM and Knn MHMM are better than uni-modal HMM (i.e. visual-only HMM and audio-only HMM). That suggests that multiple modalities (audio and visual modalities) can provide more affective information and have the potential to obtain better recognition performance than a single modality.

In Figure 4, the accuracy of Acc MHMM equals to visual-only HMM for male data but worse than visual-only HMM for female data. Ldc MHMM is worse than visual-only HMM in female and male cases. Both of Acc MHMM and Ldc MHMM are linear bimodal fusion. That suggests that the fusion method play an important role in audio-visual emotion recognition. Although the linear bimodal combination is

Table 5. Audio-visual Emotion Recognition

Bimodal Fusion		Combination scheme	Accuracy (%)
Female	Acc MHMM	Weighting	84.38
	Ldc MHMM	Training	75.00
	Knn MHMM	Training	87.50
	AdaBoost MHMM	Training	90.36
Male	Acc MHMM	Weighting	84.85
	Ldc MHMM	Training	81.82
	Knn MHMM	Training	87.88
	AdaBoost MHMM	Training	89.39

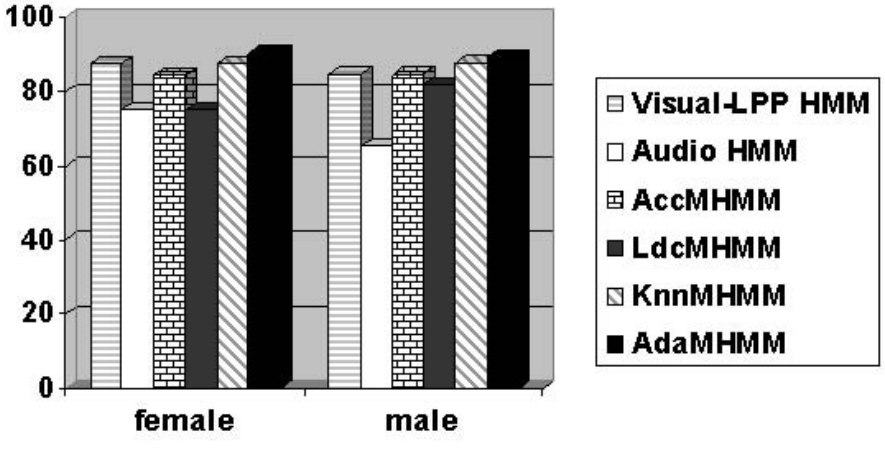


Fig. 6. Performance comparison among different modalities and different fusions

Table 6. Confusion Matrix for Female Emotion Recognition

Female	Detected		
Desired	%	Positive	Negative
	Positive	94.44	5.56
	Negative	10.87	89.13

reasonable and intuitive, it is not guaranteed to obtain the optimal combination at realistic application. Even it is possible that this combination is worse than individual component performance, as shown in our experiments.

The confusion matrixes of emotion recognition for two subjects are shown in Table 6 and 7. These results demonstrate that negative emotions are more difficult to

Table 7. Confusion Matrix for Male Emotion Recognition

Male	Detected		
Desired	%	Positive	Negative
	Positive	91.67	8.33
	Negative	13.33	86.67

recognize than positive emotions. We noticed that adult subjects tend to inhibit negative emotion expressions in this interactive interview context. Thus, the negative emotions are shorter and more subtle than positive emotions.

8 Conclusion

Emotion analysis has been attracting increased attention of researchers from various disciplines because changes in a speaker's affective states play a significant role in human communication. Most of current automatic facial expression recognition approaches are based on artificial materials of deliberately expressed emotions and uni-modal methods.

In this paper, we explore audio-visual recognition of spontaneous emotions occurring in Adult Attachment Interview (AAI) in which adults talked about past parent-child experiences. We propose an approach for this realistic application, which includes the audio-visual labeling assumption and Adaboost multi-stream hidden Markov model to integrate facial expression and prosody expression. Our preliminary experimental results from two video of about-40-minute-long AAI suggest the validation of our approach for spontaneous emotion recognition. In the future, our approach in this paper will be evaluated on more AAI data. In addition, we will explore person-independent emotion recognition in which training data and testing data are from different subjects.

Our work is based on the assumption that facial expressions are consistent of vocal expressions at the coarse emotion level (positive and negative emotions). Although this assumption is valid at most circumstances, blended emotions can occur when speakers have conflict intension [23]. The exploration of recognition of the blended emotions is our future work.

Acknowledgments. We like to thank Prof. Maja Pantic and Anton Nijholt for providing valuable comments. The work was supported in part by a Beckman Postdoctoral Fellowship and in part by National Science Foundation Grant CCF 04-26627.

References

1. Picard, R.W., Affective Computing, MIT Press, Cambridge, 1997.
2. Litman, D.J. and Forbes-Riley, K., Predicting Student Emotions in Computer-Human Tutoring Dialogues. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July 2004

3. Kapoor, A. and Picard, R.W., Multimodal Affect Recognition in Learning Environments, ACM Multimedia, 2005, 677-682
4. Viola P. 2004. Robust Real-Time Face Detection. Int. Journal of Computer Vision. 57(2), 137-154
5. Polzin, S.T. and Waibel, A. (1999), Pronunciation Variations in Emotional Speech, Proceedings of the ESCA Workshop, 103-108
6. Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C. (2005), ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance, Neural Networks, 18: 437-444
7. Steeneken, H.J.M. and Hansen, J.H.L. (1999), Speech under stress conditions: Overview of the effect of speech production and on system performance, Int. Conf. on Acoustics, Speech, and Signal Processing, 4:2079-2082
8. Okawa, S., Bocchieri, E. and Potamianos, A., Multi-band Speech Recognition in noisy environments, ICASSP, 1998, 641-644
9. Garg, A., Potamianos, G., Neti, C. & Huang, T.S., Frame-dependent multi-stream reliability indicators for audio-visual speech recognition, ICASSP, 2003.
10. Ekman P, Friesen WV, Ellsworth P. (1972). Emotion in the Human Face. Elmsford, NY: Pergamon.
11. Izard CE. 1971. The face of Emotion. New York: Appleton-Century_Crofts
12. Scherer (2004), Feelings integrate the central representation of appraisal-driven response organization in emotion. In Manstead, A.S.R., Frijda, N.H. & Fischer, A.H. (Eds.), Feelings and emotions, The Amsterdam symposium (pp. 136-157). Cambridge: Cambridge University Press, 136-157.
13. Ekman P, Friensen WV, Hager JC. 2002. Facial Action Unit System. Published by A Human Face.
14. Ekman P and Rosenberg EL. 2005. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using Facial Action Coding System. 2nd Ed. Oxford University Express.
15. Russell JA, Bachorowski JA and Fernandez-Dols JM. 2003. Facial and Vocal Expressions of Emotion. Annual Review Psychology, 2003, 54:329-49
16. Ekman P. and Oster H. 1979. Facial Expressions of Emotion. Annual Review Psychology. 30:527-54
17. Roisman, G.I., Tsai, J.L., Chiang, K.S.(2004), The Emotional Integration of Childhood Experience: Physiological, Facial Expressive, and Self-reported Emotional Response During the Adult Attachment Interview, Developmental Psychology, Vol. 40, No. 5, 776-789
18. Cohn JF and Tronick EZ. 1988. Mother Infant Interaction: the sequence of dyadic states at three, six and nine months. Development Psychology, 23, 68-77
19. Fried E. 1976. The impact of nonverbal communication of facial affect on children's learning. PhD thesis, Rutgers University, New Brunswick, NJ
20. Ekman P, Matsumoto D, and Friesen WV. 2005. Facial Expression in Affective Disorders. In What the Face Reveals. Edited by Ekman P and Rosenberg EL. 429-439
21. Zeng, Z, Fu, Y., Roisman, G.I., Wen, Z., Hu, Y. and Huang, T.S., One-class classification on spontaneous facial expressions, Automatic Face and Gesture Recognition, 281 – 286, 2006
22. Bourlard, H. and Dupont, S., A new ASR approach based on independent processing and recombination of partial frequency bands, ICSLP 1996
23. Devillers, L., Vidrascu L. and Lamel L., Challenges in real-life emotion annotation and machine learning based detection, Neural Networks, 18(2005), *407-422*

24. Ekman, P., Hager, J.C., Methvin, C.H. and Irwin, W., Ekman-Hager Facial Action Exemplars, unpublished, San Francisco: Human Interaction Laboratory, University of California
25. Kanade, T., Cohn, J., and Tian, Y. (2000), Comprehensive Database for Facial Expression Analysis, In Proceeding of International Conference on Face and Gesture Recognition, 46-53
26. Pantic, M., Valstar, M.F, Rademaker, R. and Maat, L. (2005), Web-based database for facial expression analysis, Int. Conf. on Multimedia and Expo
27. JAFFE: www.mic.atr.co.jp/~mlyons/jaffe.html
28. Chen, L.S, Joint Processing of Audio-Visual Informa-tion for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC, 2000
29. Cowie, R., Douglas-Cowie E. and Cox, C., Beyond emotion archetypes: Databases for emotion modelling using neural networks, 18(2005), 371-388
30. Ekman, P. and Rosenberg, E. (Eds.), What the face reveals. NY: Oxford University, 1997
31. Cohn, J.F. and Schmidt, K.L.(2004), The timing of Facial Motion in Posed and Spontaneous Smiles, International Journal of Wavelets, Multiresolution and Information Processing, 2, 1-12
32. Valstar MF, Pantic M, Ambadar Z and Cohn JF. 2006. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimedia Interfaces. 162-170
33. Ekman, P. (1994), Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique, Psychological Bulletin, 115(2): 268-287
34. Pantic M., Rothkrantz, L.J.M., Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept. 2003, 1370-1390
35. Pantic, M., Sebe, N., Cohn, J.F. and Huang, T., Affective Multimodal Human-Computer Interaction, in Proc. ACM Int'l Conf. on Multimedia, November 2005, 669-676
36. Sebe, N., Cohen, I., Gevers, T., and Huang, T.S. (2005), Multimodal Approaches for Emotion Recognition: A Survey, In Proc. Of SPIE-IS&T Electronic Imaging, SPIE Vol 5670: 56-67
37. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G., Emotion Recognition in Human-Computer Interaction, IEEE Signal Processing Magazine, January 2001, 32-80
38. Chen, L. and Huang, T. S., Emotional expressions in audiovisual human computer interaction, Int. Conf. on Multimedia & Expo 2000, 423-426
39. Chen, L., Huang, T. S., Miyasato, T., and Nakatsu, R., Multimodal human emotion/ expression recognition, Int. Conf. on Automatic Face & Gesture Recognition 1998, 396-401
40. De Silva, L. C., and Ng, P. C., Bimodal emotion recognition, Int. Conf. on Automatic Face & Gesture Recognition 2000, 332-335
41. Yoshitomi, Y., Kim, S., Kawano, T., and Kitazoe, T., Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in Proc. ROMAN 2000, 178-183
42. Hoch, S., Althoff, F., McGlaun, G., Rigoll, G., Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005
43. Wang, Y. and Guan, L., Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128
44. Zeng, Z., Hu, Y., Liu, M., Fu, Y. and Huang, T.S., Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition, in Proc. ACM Int'l Conf. on Multimedia, 2005, 65-68

45. Zeng, Z., Tu, J., Pianfetti , P., Liu, M., Zhang, T., et al., Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI, Int. Conf. Computer Vision and Pattern Recognition. 2005: 967-972
46. Zeng, Z., Tu., J., Liu, M., Huang, T.S. (2005), Multi-stream Confidence Analysis for Audio-Visual Affect Recognition, the Int. Conf. on Affective Computing and Intelligent Interaction, 946-971
47. Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., and Huang, T.S. (2005), Audio-visual Affect Recognition in Activation-evaluation Space, Int. Conf. on Multimedia & Expo, 828-831.
48. Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Levinson, S. (2007), Audio-visual Affect Recognition, IEEE Transactions on Multimedia, in press
49. Song, M., Bu, J., Chen, C., and Li, N., Audio-visual based emotion recognition—A new approach, Int. Conf. Computer Vision and Pattern Recognition. 2004, 1020-1025
50. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M. et al., Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. 2004. Int. Conf. Multimodal Interfaces. 205-211
51. Hoch, S., Althoff, F., McGlaun, G., Rigoll, G., Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005
52. Wang, Y. and Guan, L., Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128
53. Go HJ, Kwak KC, Lee DJ, and Chun MG. 2003. Emotion recognition from facial image and speech signal. Int. Conf. of the Society of Instrument and Control Engineers. 2890-2895
54. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J.(2005), Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, IEEE CVPR'05
55. Sebe, N., Lew, M.S., Cohen, I., Sun, Y., Gevers, T., Huang, T.S.(2004), Authentic Facial Expression Analysis, Int. Conf. on Automatic Face and Gesture Recognition
56. Zeng, Z., Fu, Y., Roisman, G.I., Wen, Z., Hu, Y., and Huang, T.S. (2006). Spontaneous Emotional Facial Expression Detection. Journal of Multimedia, 1(5): 1-8.
57. Valstar MF, Pantic M and Ambadar Z, and Cohn JF. 2006. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimodal Interfaces. 162-170
58. Cohn JF, Reed LI, Ambadar Z, Xiao J, and Moriyama T. 2004. Automatic Analysis and recognition of brow actions and head motion in spontaneous facial behavior. Int. Conf. on Systems, Man & Cybernetics, 1, 610-616
59. Litman, D.J. and Forbes-Riley, K., Predicting Student Emotions in Computer-Human Tutoring Dialogues. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July 2004
60. Batliner A, Fischer K, Hubera R, Spilkera J and Noth E. 2003. How to find trouble in communication. Speech Communication, Vol. 40, 117-143.
61. Neiberg D, Elenius K, and Laskowski K. 2006. Emotion Recognition in Spontaneous Speech Using GMM. Int. Conf. on Spoken Language Processing, 809-812
62. Ang J, Dhillon R, Krupski A, et al. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog, ICSLP.
63. Fragopanagos, F. and Taylor, J.G., Emotion recognition in human-computer interaction, Neural Networks, 18 (2005) 389-405
64. Garidakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A. and Karpouzis, K. 2006. Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition. Int. Conf. on Multimodal Interfaces. 146-154

65. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'Feeltrace': an instrument for recording perceived emotion in real time. Proceedings of the ISCA Workshop on Speech and Emotion, 19–24
66. Maat L and Pantic M. 2006. Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios. Int. Conf. on Multimodal Interfaces. 171-178
67. Lanitis, A., Taylor, C. and Cootes, T. (1995), A Unified Approach to Coding and Interpreting Face Images, in Proc. International Conf. on Computer Vision, 368-373
68. Black, M. and Yacoob, Y.(1995), Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motion, in Proc. Int. Conf. on Computer Vision, 374-381
69. Rosenblum, M., Yacoob, Y., and Davis, L. (1996), Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture, IEEE Trans. On Neural Network, 7(5):1121-1138
70. Essa, I. and Pentland, A. (1997), Coding, Analysis, Interpretation, and Recognition of Facial Expressions, IEEE Trans. On Pattern Analysis and Machine Intelligence, 19(7): 757-767
71. Cohen, L., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003), Facial expression recognition from video sequences: Temporal and static modeling, Computer Vision and Image Understanding, 91(1-2):160-187
72. Tian, Y., Kanade, T., Cohn, J.F. (2001), Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2): 97-115
73. Pantic M and Patras I. 2006. 'Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences', IEEE Transactions on Systems, Man and Cybernetics - Part B, vol. 36, no. 2, pp. 433-449
74. Kwon, O.W., Chan, K., Hao, J., Lee, T.W (2003), Emotion Recognition by Speech Signals, EUROSPEECH.
75. Polzin, Thomas (1999), Detecting Verbal and Non-verbal cues in the communication of emotion, PhD thesis, Carnegie Mellon University
76. Amir, N. and Ron, S. (1998), Toward Automatic Classification of Emotions in Speech, in Proc. ICSLP, 555-558
77. Dellaert, F., Polzin, T., and Waibel, A. (1996), Recognizing Emotion in Speech, In Proc. ICSLP, 1970-1973
78. Petrushin, V.A. (2000), Emotion Recognition in Speech Signal, In Proc. ICSLP, 222-225
79. Pantic M, Pentland A, Nijholt A and Huang TS. 2006. Human Computing and Machine Understanding of Human Behavior: A Survey. Int. Conf. Multimodal Interfaces. 233-238
80. Huang, D. (1999), Physiological, subjective, and behavioral Responses of Chinese American and European Americans during moments of peak emotional intensity, honor Bachelor thesis, Psychology, University of Minnesota.
81. Tao, H. and Huang, T.S., Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode, IEEE CVPR'99, vol.1, pp. 611-617, 1999
82. Wen Z and Huang T. 2003. Capturing Subtle Facial Motions in 3D Face Tracking. Intl. Conf. on Computer Vision (ICCV). 1343-1350
83. He, X., Yan, S., Hu, Y., and Zhang, H, Learning a Locality Preserving Subspace for Visual Recognition, Int. Conf. on Computer Vision, 2003

Modeling Naturalistic Affective States Via Facial, Vocal, and Bodily Expressions Recognition

Kostas Karpouzis[1], George Caridakis[1], Loic Kessous[2], Noam Amir[2], Amaryllis Raouzaiou[1], Lori Malatesta[1], and Stefanos Kollias[1]

[1] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Politechnioupoli, Zographou, Greece
{kkarpou, gcari, araouz, lori, stefanos}@image.ntua.gr
[2] Tel Aviv Academic College of Engineering
218 Bnei Efraim St. 69107, Tel Aviv, Israel
{kessous, noama}@post.tau.ac.il

Abstract. Affective and human-centered computing have attracted a lot of attention during the past years, mainly due to the abundance of devices and environments able to exploit multimodal input from the part of the users and adapt their functionality to their preferences or individual habits. In the quest to receive feedback from the users in an unobtrusive manner, the combination of facial and hand gestures with prosody information allows us to infer the users' emotional state, relying on the best performing modality in cases where one modality suffers from noise or bad sensing conditions. In this paper, we describe a multi-cue, dynamic approach to detect emotion in naturalistic video sequences. Contrary to strictly controlled recording conditions of audiovisual material, the proposed approach focuses on sequences taken from nearly real world situations. Recognition is performed via a 'Simple Recurrent Network' which lends itself well to modeling dynamic events in both user's facial expressions and speech. Moreover this approach differs from existing work in that it models user expressivity using a dimensional representation of activation and valence, instead of detecting discrete 'universal emotions', which are scarce in everyday human-machine interaction. The algorithm is deployed on an audiovisual database which was recorded simulating human-human discourse and, therefore, contains less extreme expressivity and subtle variations of a number of emotion labels.

Keywords: Affective interaction, multimodal analysis, facial expressions, prosody, hand gestures, neural networks.

1 Introduction

The introduction of the term 'affective computing' by R. Picard [45] epitomizes the fact that computing is no longer considered a 'number crunching' discipline, but should be thought of as an interfacing means between humans and machines and sometimes even between humans alone. To achieve this, application design must take into account the ability of humans to provide multimodal input to computers, thus moving away from the monolithic window-mouse-pointer interface paradigm and

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 91–112, 2007.

utilizing more intuitive concepts, closer to human niches ([1], [4]). A large part of this naturalistic interaction concept is expressivity [46], both in terms of interpreting the reaction of the user to a particular event or taking into account their emotional state and adapting presentation to it, since it alleviates the learning curve for conventional interfaces and makes less technology-savvy users feel more comfortable. In this framework, both speech and facial expressions are of great importance, since they usually provide a comprehensible view of users' reactions.

The complexity of the problem relies in the combination of the information extracted from modalities, the interpretation of the data through time and the noise alleviation from the natural setting. The current work aims to interpret sequences of events thus modeling the user's behavior through time. With the use of a recurrent neural network, the short term memory, provided through its feedback connection, works as a memory buffer and the information remembered is taken under consideration in every next time cycle. Theory on this kind of network back up the claim that it is suitable for learning to recognize and generate temporal patterns as well as spatial ones [20].

The naturalistic data chosen as input is closer to human reality since the dialogues are not acted, and the expressivity is not guided by directives (e.g. Neutral expression → one of the six universal emotions → neutral). This amplifies the difficulty in discerning facial expressions and speech patterns. Nevertheless it provides the perfect test-bed for the combination of the conclusions drawn from each modality in one time unit and use as input in the following sequence of audio and visual events analyzed.

In the area of unimodal emotion recognition there have been many studies using different, but single, modalities. Facial expressions [21], [50], [31], vocal features [41], [24] and physiological signals [48] have been used as inputs during these attempts, while multimodal emotion recognition is currently gaining ground ([26], [28], [32], [53]).

A wide variety of machine learning techniques have been used in emotion recognition approaches ([50], [31], [42]). Especially in the multimodal case [34], they all employ a large number of audio, visual or physiological features, a fact which usually impedes the training process; therefore, we need to find a way to reduce the number of utilized features by picking out only those related to emotion. An obvious choice for this is neural networks, since they enable us to pinpoint the most relevant features with respect to the output, usually by observing their weights. Although such architectures have been successfully used to solve problems that require the computation of a static function, where output depends only upon the current input, and not on any previous inputs, this is not the case in the domain of emotion recognition. One of the reasons for this is that expressivity is a dynamic, time-varying concept, where it is not always possible to deduce an emotional state merely by looking at a still image. As a result, Bayesian approaches which lend themselves nicely to similar problems [37], need to be extended to include support for time-varying features. Picard [46] and Cohen [16] propose the use of Hidden Markov Models (HMMs) to model discrete emotional states (interest, joy or distress) and use them to predict the probability of each one, given a video of a user. However, this process needs to build a single HMM for each of the examined cases (e.g. each of the universal emotions), making it more suitable in cases where discrete emotions need to be estimated. In our case building dedicated HMMs for each of the quadrants of the

emotion representation would not perform well, since each of them contains emotions expressed with highly varying features (e.g. anger and fear both lie in the negative/active quadrant) which cannot be modeled using a single model.

A more suitable choice would be RNNs (Recurrent Neural Networks) where past inputs influence the processing of future inputs [35]. RNNs possess the nice feature of modeling explicitly time and memory ([49], [14], [9]), catering for the fact that emotional states are not fluctuating strongly, given a short period of time. Additionally, they can model emotional transitions and not only static emotional representations, providing a solution for diverse feature variation and not merely for neutral to expressive and back to neutral, as would be the case for HMMs.

2 Induction of Natural Emotions – Data Collection

Research on signs of emotion emerged as a technical field around 1975, with research by Ekman and his colleagues [38] on encoding emotion-related features of facial expression, and by Williams and Stevens [52] on emotion in the voice. The early paradigms simplified their task by concentrating on emotional extremes – often simulated, and not always by skilled actors. Most of the data used in research on speech and emotion has three characteristics: the emotion in it is simulated by an actor (not necessarily trained); the actor is reading preset material; and he or she is aiming to simulate full-blown emotion.

That kind of material has obvious attractions: it is easy to obtain, and it lends itself to controlled studies. However, it has become reasonably clear that it does not do a great deal to illuminate the way face and speech express emotion in natural settings. The 1990's saw growing interest in naturalistic data, but retained a focus on cases where emotion was at or approaching an extreme. The major alternative is to develop techniques which might be called directed elicitation – techniques designed to induce states that are both genuinely emotional and likely to involve speech.

Most of these tasks have a restricted range. They provide more control and higher data rates than other methods, but they still tend to elicit weak negative emotions, and they often impose constraints on the linguistic form and content of the speech which may restrict generalization. One of them, SAL, was used to acquire the data processed by the presented system.

The SAL scenario [42] is a development of the ELIZA concept introduced by Weizenbaum [19]. The user communicates with a system whose responses give the impression of sympathetic understanding, and that allows a sustained interaction to build up. In fact, though, the system takes no account of the user's meaning: it simply picks from a repertoire of stock responses on the basis of surface cues extracted from the user's contributions. A second factor in the selection is that the user selects one of four 'artificial listeners' to interact with at any given time. Each one will try to initiate discussion by providing cues mapped to each of the four quadrants defined by valence and activation – 'Spike' is provocative or angry (negative/active), while 'Poppy' is always happy (positive/active). SAL took its present form as a result of a good deal of pilot work [10]. In that form, it provides a framework within which users do express a considerable range of emotions in ways that are virtually unconstrained. The process depends on users' co-operation – they are told that it is like an emotional gym, and

they have to use the machinery it provides to exercise their emotions. But if they do enter into the spirit, they can move through a very considerable emotional range in a recording session or a series of recording sessions: the 'artificial listeners' are designed to let them do exactly that.

As far as emotion representation is concerned we use the Activation-Evaluation space. Activation-Evaluation space as a representation has great appeal as it is both simple, while at the same time makes it possible to capture a wide range of significant issues in emotion. The concept is based on a simplified treatment of two key themes:

- *Valence:* The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, i.e., they are centrally concerned with positive or negative evaluations of people or things or events.
- *Activation level:* Research from Darwin forward has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e., the strength of the person's disposition to take some action rather than none.

Dimensional representations are attractive mainly because they provide a way of describing emotional states that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where a wide range of emotional states occur. Similarly, they are much more able to deal with non-discrete emotions and variations in emotional state over time, since cases of changing from one universal emotion label to another would not make much sense in real life scenarios.

The available SAL Data Set is given along with some of their important features in Table 1, while Figure 1 shows some frames of the processed naturalistic data.

Table 1. Summary of the processed SAL data

Data Set	
Subjects	2 males, 2 females
Passages	76
Tunes	~1600
Emotion space coverage	Yes, all quadrants
FeelTrace ratings	4 rators
Transcripts	Yes
Text Post-Processing	No

Naturalistic data goes beyond extreme emotions, as is usually the case in existing approaches (see Fig. 1 for examples of expressive images, not so easy to classify using universal emotions), and concentrates on more natural emotional episodes that happen more frequently in everyday discourse.

Fig. 1. Frames from the SAL dataset

3 Extraction of Facial Features

An overview of the proposed methodology is illustrated in Fig. 2. The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, i.e. eyes, eyebrows, mouth and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria [18] to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation.

Since this procedure essentially locates and tracks points in the facial area, we chose to work with MPEG-4 FAPs and not Action Units (AUs), since the former are explicitly defined to measure the deformation of these feature points. In addition to this, discrete points are easier to track in cases of extreme rotations and their position

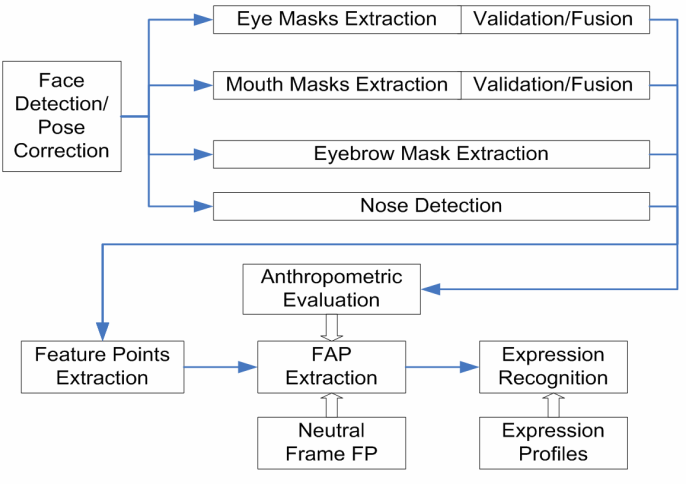


Fig. 2. High-level overview of the facial feature extraction process

can be estimated based on anthropometry in cases of occlusion, whereas this is not usually the case with whole facial features. Another feature of FAPs which proved useful is their value (or magnitude) which is crucial in order to differentiate cases of varying activation of the same emotion (e.g. joy and exhilaration) [5] and exploit fuzziness in rule-based systems [50]. Measurement of Facial Animation Parameters (FAPs) requires the availability of a frame where the subject's expression is found to be neutral. This frame will be called the *neutral frame* and is manually selected from video sequences to be analyzed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks are used to extract 19 Feature Points (FPs) [5]**;** Feature Points obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the Facial Animation Parameters (FAPs). Confidence levels on FAP estimation are derived from the equivalent feature point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

3.1 Face Detection and Pose Estimation

In the proposed approach facial features including eyebrows, eyes, mouth and nose are first detected and localized. Thus, a first processing step of face detection and pose estimation is carried out as described below, to be followed by the actual facial feature extraction process described in the following section. At this stage, it is assumed that an image of the user at neutral expression is available, either a-priori, or captured before interaction with the proposed system starts.

The goal of face detection is to determine whether or not there are faces in the image, and if yes, return the image location and extent of each face [30]. Face detection can be performed with a variety of methods. In this paper we used nonparametric discriminant analysis with a *Support Vector Machine* (SVM) which classifies face and non-face areas reducing the training problem dimension to a fraction of the original with negligible loss of classification performance [44],[10].

(a) (b)

Fig. 3. Feature-Candidate areas: (a) full frame (352x288). (b) zoomed (90x125).

The face detection step provides a rectangle head boundary which includes all facial features as shown in Fig. 3. The latter can be then segmented roughly using static anthropometric rules into three overlapping rectangle regions of interest which include both facial features and facial background; these three *feature-candidate areas* include the left eye/eyebrow, the right eye/eyebrow and the mouth. In the following we utilize these areas to initialize the feature extraction process. Scaling does not affect feature-candidate area detection, since the latter is proportional to the head boundary extent, extracted by the face detector.

The accuracy of feature extraction depends on head pose. In this paper we are mainly concerned with roll rotation, since it is the most frequent rotation encountered in real life video sequences. Small head yaw and pitch rotations which do not lead to feature occlusion do not have a significant impact on facial expression recognition. The face detection techniques described in the former section is able to cope with head roll rotations up to 30^o. This is a quite satisfactory range in which the feature-candidate areas are large enough so that the eyes reside in the eye-candidate search areas defined by the initial segmentation of a rotated face.

To estimate the head pose we first locate the left and right eyes in the detected corresponding eye candidate areas. After locating the eyes, we can estimate head roll rotation by calculating the angle between the horizontal plane and the line defined by the eye centers. To increase speed and reduce memory requirements, the eyes are not detected on every frame using the neural network. Instead, after the eyes are located in the first frame, two square grayscale eye templates are created, containing each of the eyes and a small area around them. The size of the templates is half the eye-center distance (bipupil breadth, D_{bp}). For the following frames, the eyes are located inside the two eye-candidate areas, using template matching which is performed by finding the location where the sum of absolute differences (SAD) is minimized.

After head pose is computed, the head is rotated to an upright position and new feature-candidate segmentation is performed on the head using the same rules so as to ensure facial features reside inside their respective candidate regions. These regions containing the facial features are used as input for the facial feature extraction stage, described in the following section.

3.2 Automatic Facial Feature Detection and Boundary Extraction

To be able to compute MPEG-4 FAPs, precise feature boundaries for the eyes, eyebrows and mouth have to be extracted. Eye boundary detection is usually performed by detecting the special color characteristics of the eye area [47], by using luminance projections, reverse skin probabilities or eye model fitting. Mouth boundary detection in the case of a closed mouth is a relatively easily accomplished task. In case of an open mouth, several methods have been proposed which make use of intensity or color information. Color estimation is very sensitive to environmental conditions, such as lighting or capturing camera's characteristics and precision. Model fitting usually depends on ellipse or circle fitting, using Hough-like voting or corner detection [23]. Those techniques while providing accurate results in high resolution images, are unable to perform well with low video resolution which lack high frequency properties; such properties which are essential for efficient corner detection and feature border trackability [8], are usually lost due to analogue video media transcoding or low quality digital video compression.

In this work, nose detection and eyebrow mask extraction are performed in a single stage, while for eyes and mouth which are more difficult to handle, multiple (four in our case) masks are created taking advantage of our knowledge about different properties of the feature area; the latter are then combined to provide the final estimates as shown in Fig. 4. More technical details can be found at [50].

3.2.1 Eye Boundary Detection

Luminance and color information fusion mask tries to refine eye boundaries extracted by the neural network described earlier building on the fact that eyelids usually appear darker than skin due to eyelashes and are almost always adjacent to the iris. The initial mask provided by the neural network is thresholded and the distance transformation of the resulting mask gives as the first eye mask.

This second approach is based on eyelid edge detection. Eyelids reside above and below the eye centre, which has already been estimated by the neural network. Taking advantage of their mainly horizontal orientation, eyelids are easily located through edge detection. By combining the canny edge detector and the vertical gradient we are locating the eyelids and the space between them is considered the eye mask.

A third mask is created for each of the eyes to strengthen the final mask fusion stage. This mask is created using a region growing technique; the latter usually gives very good segmentation results corresponding well to the observed edges. Construction of this mask relies on the fact that facial texture is more complex and darker inside the eye area and especially in the eyelid-sclera-iris borders, than in the areas around them. Instead of using an edge density criterion, we developed a simple but effective iterative method to estimate both the eye centre and eye mask based on the standard deviation of the luminance channel.

Finally, a second luminance-based mask is constructed for eye/eyelid border extraction. In this mask, we compute the normal luminance probability resembling to the mean luminance value of eye area defined by the NN mask. From the resulting

(a) (b) (c)

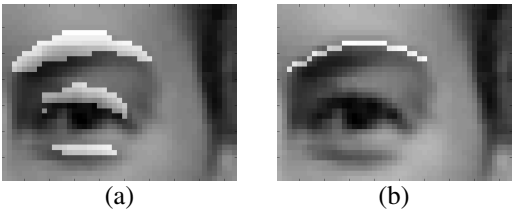
Fig. 4. Eye masks

probability mask, the areas with a given confidence interval are selected and small gaps are closed with morphological filtering. The result is usually a blob depicting the boundaries of the eye. In some cases, the luminance values around the eye are very low due to shadows from the eyebrows and the upper part of the nose. To improve the outcome in such cases, the detected blob is cut vertically at its thinnest points from both sides of the eye centre; the resulting mask's convex hull is used as the Luminance mask (Figure 4).

3.2.2 Eyebrow Boundary Detection

Eyebrows are extracted based on the fact that they have a simple directional shape and that they are located on the forehead, which due to its protrusion, has a mostly uniform illumination. Each of the left and right eye and eyebrow-candidate images shown in Figure 3 is used for brow mask construction.

The first step in eyebrow detection is the construction of an edge map of the grayscale eye/eyebrow-candidate image. This map is constructed by subtracting the dilation and erosion of the grayscale image using a line structuring element of size n and then thresholding the result. The selected edge detection mechanism is appropriate for eyebrows because it can be directional; it preserves the feature's original size and can be combined with a threshold to remove smaller skin anomalies such as wrinkles. The above procedure can be considered as a non-linear high-pass filter.



(a) (b)

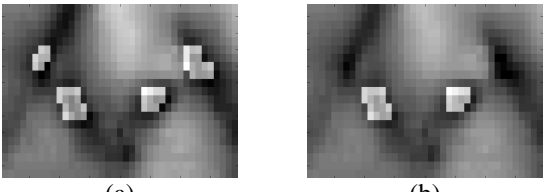
Fig. 5. (a) eyebrow-candidates. (b) selected eyebrow mask.

Each connected component on the edge map is labeled and then tested against a set of filtering criteria. These criteria were formed through statistical analysis of the eyebrow lengths and positions on 20 persons of the ERMIS SAL database [10]. Firstly, the major axis is found for each component through principal component analysis (PCA). All components whose major axis has an angle of more than 30 degrees with the horizontal plane are removed from the set. From the remaining components, those whose axis length is smaller than a given threshold are removed.

Finally components with a lateral distance from the eye centre greater than a threshold calculated by anthropometric criteria are removed and the top-most remaining is selected resulting in the eyebrow mask (see Fig. 5).

3.2.3 Nose Localization

The nose is not used for expression estimation by itself, but is a fixed point that facilitates distance measurements for FAP estimation, thus, its boundaries do not have to be precisely located. Nose localization is a feature frequently used for face tracking and usually based on nostril localization; nostrils are easily detected based on their low intensity. Since inter-ocular distance in our images in the order of 50 pixels, nostril motion is limited, making them almost fixed and a good choice for a reference point.



(a) (b)

Fig. 6. (a) nostril candidates. (b) selected nostrils.

The facial area above the mouth-candidate components area is used for nose location. The respective luminance image is thresholded and connected objects of the derived binary map are labeled. In bad lighting conditions, long shadows may exist along either side of the nose. For this reason, anthropometric data about the distance of left and right eyes (bipupil breadth, etc.) is used to reduce the number of candidate objects. This has proven to be an effective way to remove most outliers without causing false negative results while generating the nostril mask shown in Figure 6a.

Horizontal nose coordinate is predicted from the coordinates of the two eyes. Each of the connected component horizontal distances from the predicted nose centre is compared to the average inter-nostril distance and components with the largest ones are considered as outliers. Those who qualify enter two separate lists, one including left-nostril candidates and one with right-nostril candidates based on their proximity to the left or right eye. Those lists are sorted according to their luminance and the two objects with the lowest values are retained from each list. The largest object is finally kept from each list and labeled as the left and right nostril respectively, as shown in Figure 6b. The nose centre is defined as the midpoint of the nostrils.

3.2.4 Mouth Detection

At first, mouth boundary extraction is performed on the mouth-candidate facial area depicted in Figure 3. An MLP neural network is trained to identify the mouth region using the neutral image. Since the mouth is closed in the neutral image, a long low-luminance region exists between the lips. The initial mouth-candidate luminance image is simplified to reduce the presence of noise, remove redundant information and produce a smooth image that consists mostly of flat and large regions of interest.

Alternating Sequential Filtering by Reconstruction (ASFR) is thus performed on the initial mouth mask to produce a filtered image. ASFR ensures preservation of object boundaries through the use of connected operators [25]. The major axis of each connected component is computed through PCA analysis, and the one with the longest axis is selected. The latter is subsequently dilated vertically and this procedure results in a mask which includes the lips. The neural network trained on the neutral-expression frame, is then used on other frames to produce an estimate of the mouth area: neural network output on the mouth-candidate image is thresholded and those areas with high confidence are kept, to form a binary map containing several small sub-areas. The convex hull of these areas is calculated to generate the first final mask.

The second approach which produces a generic edge connection mask, the mouth luminance channel is again filtered using ASFR for image simplification. The horizontal morphological gradient of the original mouth mask is calculated similarly to the eyebrow binary edge map detection resulting in an intermediate mask. Since the nose has already been detected, its vertical position is known. The connected elements of the intermediate mask are labeled and those too close to the nose are removed. From the rest of the map, very small objects are removed by thresholding. Morphological closing is then performed and the longest of the remaining objects in horizontal sense is selected as the second mouth mask.

The problem of most intensity-based methods, that try to estimate mouth opening, is existence of upper teeth, i.e., those appearing between the upper and lower lip altering saturation and intensity uniformity. A new method is proposed next to cope with this problem. First, the mouth-candidate luminance channel is thresholded using a low threshold providing an estimate of the mouth interior area, or the area between the lips in case of a closed mouth. The threshold used is estimated adaptively.

In the resulting binary map, all connected objects adjacent to the border are removed. We now examine two cases separately: either we have no apparent teeth and the mouth area is denoted by a cohesive dark area (case 1) or teeth are apparent and thus two dark areas appear at both sides of the teeth (case 2). It should be noted that those areas appear even in large extensive smiles. The largest connected object is then selected and its centroid is found. If the horizontal position of its centroid is near the horizontal nose position case 1 is selected, otherwise case 2 is assumed to occur and two dark areas appear at both sides of the teeth. The two cases are quite distinguishable through this process. In case 2, the second largest connected object is also selected. A new binary map is created containing either one object in case 1 or both objects in case 2; the convex hull of this map is then calculated.

The detected lip corners provide a robust estimation of mouth horizontal extent but are not adequate to detect mouth opening. Therefore, the latter binary mask is expanded to include the lower lips. An edge map is created as follows: the mouth image gradient is calculated in the horizontal direction, and is thresholded by the median of its positive values. This mask contains objects close to the lower middle part of the mouth, which are sometimes missed because of the lower teeth. The two masks have to be combined to a final mask. An effective way of achieving this is to keep from both masks objects which are close to each other.

3.3 Final Masks Generation and Confidence Estimation

Each facial feature's masks must be fused together to produce a final mask for that feature. The most common problems, especially encountered in low quality input images, include connection with other feature boundaries or mask dislocation due to noise. In some cases some masks may have completely missed their goal and provide a completely invalid result. Outliers such as illumination changes and compression artifacts cannot be predicted and so individual masks have to be re-evaluated and combined on each new frame.

The proposed algorithms presented in the previous sections produce a mask for each eyebrow, nose coordinates, four intermediate mask estimates for each eye and three intermediate mouth mask estimates. The four masks for each eye and three mouth masks must be fused to produce a final mask for each feature. Since validation can only be done on the end result of each intermediate mask, we unfortunately cannot give different parts of each intermediate mask different confidence values, so each pixel of those masks will share the same value (see Fig. 7). We propose validation through testing against a set of anthropometric conformity criteria. Since, however some of these criteria relate either to aesthetics or to transient feature properties, we cannot apply strict anthropometric judgment.

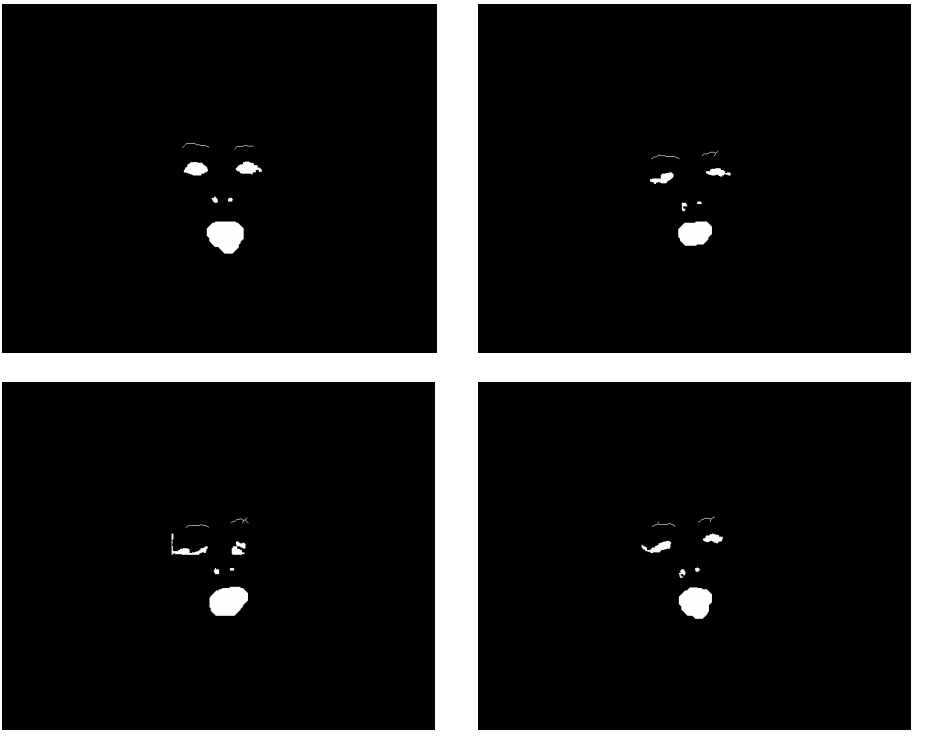


Fig. 7. Final masks for the frames shown in Fig. 1

For each mask of every feature, we employ a set of validation measurements, which are then combined to a final validation tag for that mask. Each measurement produces a validation estimate value depending on how close it is to the usually expected feature shape and position, in the neutral expression. Expected values for these measurements are defined from anthropometry data [18] and from images extracted from video sequences of 20 persons in our database [10]. Thus, a validation tag between [0,1] is attached to each mask, with higher values denoting proximity to the most expected measurement values. We want masks with very low validation tags to be discarded from the fusion process and thus those are also prevented from contribution on final validation tags.

3.4 From FP to FAP Estimation

A 25-dimensional distance vector is created containing vertical and horizontal distances between 19 extracted FPs, as shown in Fig. 8. Distances are normalized using scale-invariant MPEG-4 units, i.e. ENS, MNS, MW, IRISD and ES 36. Unit bases are measured directly from FP distances on the neutral image; for example ES is calculated as $|FP_9,FP_{13}|$.

The distance vector is created once for the neutral-expression image and for each of the subsequent frames FAPs are calculated by comparing them with the neutral frame. The value of each FAP is calculated from a set of geometric rules based on variations of distances from immovable points on the face. For example, the inner eyebrow FAPs are calculated by projecting vertically the distance of the inner eye corners, points 8 and 12 in Figure 8, to points 3 and 6 and comparing it to their distance in the neutral frame. A more detailed discussion on this procedure is found at [5].

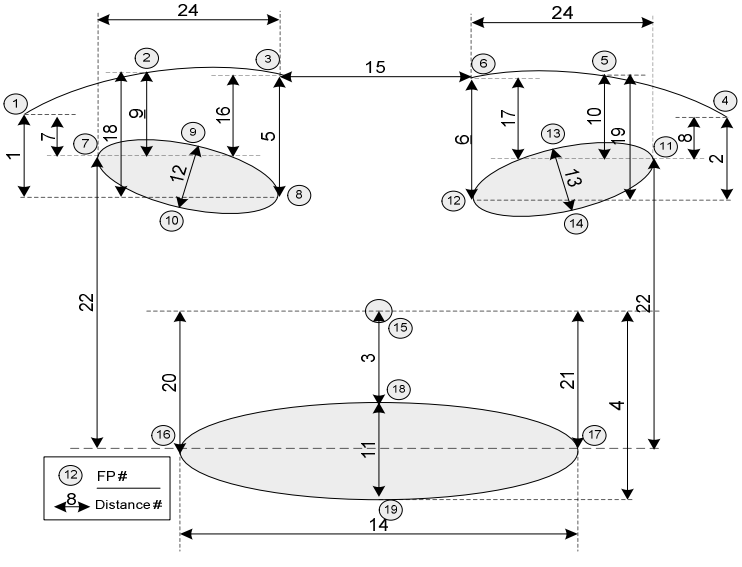


Fig. 8. Feature Point Distances

4 Hand Gesture Analysis

In order to extract emotion-related features through hand movement, we implemented a lightweight hand-tracking subsystem [22, 51]. The purpose of this subsystem was not gesture recognition per se, but the extraction of quantitative parameters related to expressivity [12, 7]; as a result, emphasis is on detailed tracking, e.g. in the case of occlusion, but on quick calculation of the position of the hands. The low-level part of the process involves the creation of *moving skin masks*, namely skin color areas which are tracked between subsequent frames. By tracking the centroid of those skin masks we produce an estimate of the user's movements, exploiting also the fact that in the

Fig. 9. Initial color mask created with skin detection

Fig. 10. Initial motion mask after pixel difference thresholded to 10% of maximum

Fig. 11. Detected moving hand segments after morphological reconstruction

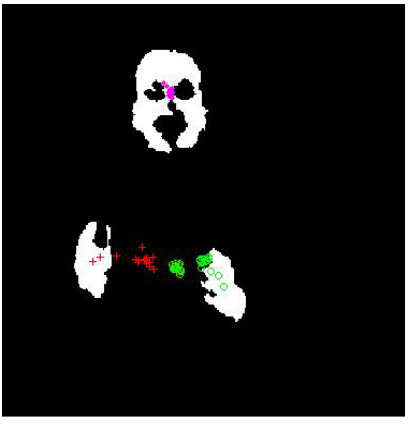


Fig. 12. Tracking of both hand objects in the "clapping" sequence

context is MMI applications, one expects to locate the head in the middle area of upper half of the frame and the hand segments near the respective lower corners.

For each given frame, a skin color probability matrix is computed by calculating the joint thresholded probability of the Cr/Cb image values (Fig. 9). Candidate areas for the 'possible-motion mask' are found by thresholding the difference pixels between the current frame and the next, resulting to Fig. 10; in essence, this mask is used to constrain processing only to moving, hence expressive, parts of the image. In the following, morphological filtering is employed on both masks in order to eliminate artifacts created by noise and background objects with color similar to the skin. The final moving skin mask is then created by fusing the processed skin and motion masks through the morphological reconstruction of the color mask using the motion mask as marker. The result of this process, after excluding the head object is shown in Fig. 11. The moving skin mask consists of many large connected areas. For the next frame a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In Fig. 12, red markers represent the position of the centroid of the detected right hand of the user, while green markers correspond to the left hand. In the case of hand object merging and splitting, e.g. in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand.

The results of this first step are utilized to calculate the gesture expressivity parameters [7]. Table 2 shows the effect that each expressivity parameter has on the production of facial expressions and hand gestures. The Spatial Extent (SPC) parameter modulates the amplitude of the movement of arms, wrists (involved in the animation of a gesture), head and eyebrows (involved in the animation of a facial expression); it influences how wide or narrow their displacement will be during the final animation. For example let us consider the eyebrows rising in the expression of surprise: if the value of the Spatial Extent parameter is very high the final position of the eyebrows will be very high in the forehead (i.e. the eyebrows move under a strong of muscular contraction). The Temporal Extent (TMP) parameter shortens or lengthens the motion of the preparation and retraction phases of the gesture as well as the onset and offset duration for facial expression. On of the effect on the face is to speed up or slow down the rising/lowering of the eyebrows. The agent animation is generated by defining some key frames and computing the interpolation curves passing through these frames. The Fluidity (FLT) and Power (PWR) parameters act on the interpolation curves. Fluidity increases/reduces the continuity of the curves allowing the system to generate more/less smooth animations. Let us consider its effect on the head: if the value of the Fluidity parameter is very low the resulting curve of the head movement will appear as generated through linear interpolation. Thus, during its final animation the head will have a jerky movement. Power introduces a gesture/expression overshooting, that is a little lapse of time in which the body part involved by the gesture reaches a point in space further than the final one. For example the frown displayed in the expression of anger will be stronger for a short period of time, and then the eyebrows will reach the final position. The last parameter, Repetition (REP), exerts an influence on gestures and head movements. It increases the number of stroke of gestures to obtain repetition of the gestures

Table 2. Expressivity parameters and effect on low-level features

	HEAD	**FACIAL EXPRESSION**	**GESTURE**
SPC	wider/narrower movement	increased/decreased muscular contraction	wider/narrower movement
TMP	shorter/longer movement speed	shorter/longer onset and offset	shorter/longer speed of preparation and retraction phases
FLT	increases/reduces continuity of head movement	increases/reduces continuity of muscular contraction	increases/reduces continuity between consecutive gestures
PWR	higher/shorter head overshooting	higher/shorter muscular contraction overshooting	more/less stroke acceleration
REP	more/less number of nods and shakes	not implemented yet	more/less number of repetitions of the same stroke

themselves in the final animation. Let us consider the gesture "wrists going up and down in front of the body with open hands and palms up", a high value of the Repetition parameter will increase the number of the up and down movements. On the other hand this parameter decreases the time period of head nods and head shakes to obtain more nods and shakes in the same lapse of time.

5 Extraction of Acoustic Features

The features used in this work are exclusively based on prosody and related to pitch and rhythm. All information related to emotion that one can extract from pitch is probably not only in these features, but the motivation of this approach is in the desire to develop and use a high level of speech prosody analysis, calculate as many features as possible and then reduce them to those uncorrelated with each other and relevant to expressivity ([42], [40]).

We analyzed each tune with a method employing prosodic representation based on perception called ‘Prosogram’ [39]. Prosogram is based on a stylization of the fundamental frequency data (contour) for vocalic (or syllabic) nuclei. It gives globally for each voiced nucleus a pitch and a length. According to a 'glissando treshold' in some cases we don’t get a fixed pitch but one or more lines to define the evolution of pitch for this nucleus. This representation is in a way similar to the 'piano roll' representation used in music sequencers. This method, based on the Praat environment, offers the possibility of automatic segmentation based both on voiced part and energy maxima. From this model - representation stylization we extracted several types of features: pitch interval based features, nucleus length features and distances between nuclei.

In musical theory, ordered pitch interval is the distance in semitones between two pitches upwards or downwards. For instance, the interval from C to G upward is 7,

but the interval from G to C downwards is −7. Using integer notation (and eventually modulo 12) ordered pitch interval, *ip*, may be defined, for any two pitches *x* and *y*, as:

$$\begin{aligned} ip\left\langle y, x\right\rangle &= x - y \\ ip\left\langle x, y\right\rangle &= y - x \end{aligned} \tag{1}$$

In this study we considered pitch intervals between successive voiced nuclei. For any two pitches *x* and *y*, where *x* precedes *y*, we calculate the interval $ip\left\langle x, y\right\rangle = y - x$, then deduce the following features.

For each tune, feature (f1) is the minimum of all the successive intervals in the tune. In a similar way, we extract the maximum (f2), the range (absolute difference between minimum and maximum) (f3), of all the successive intervals in each tune. Using the same measure, we also deduce the number of positive intervals (f4) and the number of negative intervals (f5). Using the absolute value, a measure equivalent to the unordered pitch interval in music theory, we deduce a series of similar features: minimum (f6), maximum (f7), mean (f8) and range (f9) of the pitch interval. Another series of features is also deduced from the ratio between successive intervals, here again maximum (f10), minimum (f11), mean (f12) and range (f13) of these ratios give the related features. In addition to the aforementioned features, the usual pitch features have also been used such as fundamental frequency minimum (f14), maximum (f15), mean (f16) and range (f17). The global slope of the pitch curve (f18), using linear regression, has also been added.

As was previously said, each segment (voiced 'nucleus' if it is voiced) of this representation has a length, and this has also been used in each tune to extract features related to rhythm. These features are, as previously, maximum (f19), minimum (f20), mean (f21) and range (f22). Distances between segments have also been used as features and the four last features we used are maximum (f23), minimum (f24), mean (f25) and range (f26) of these distances.

6 Fusion of Visual and Acoustic Features

As primary material we consider the audiovisual content collected using the SAL approach. This material was labeled using FeelTrace [43] by four labelers. The activation-valence coordinates from the four labelers were initially clustered into quadrants [42] and were then statistically processed so that a majority decision could be obtained about the unique emotion describing the given moment. The corpus under investigation was segmented into 1000 tunes of varying length. For every tune, the input vector consisted of the FAPs and expressivity parameters produced by the processing of the frames of the tune plus one value per SBPF (Segment Based Prosodic Feature) per tune. The fusion was performed on a frame basis, meaning that the values of the SBPFs were repeated for every frame of the tune. This approach was preferred because it preserved the maximum of the available information since SBPFs are only meaningful for a certain time period and cannot be calculated per frame.

The implementation of a RNN we used was based on an Elman network [20], [35]. The input vectors were formed as described earlier and the output classes were 4 (3 for the possible emotion quadrants, since the data for the positive/passive quadrant was negligible, and one for neutral affective state) resulting in a dataset consisting of around 13,000 records. The classification efficiency for the unimodal case was measured at 67% (visual) and 73% (prosody), but combining all modalities increased mean recognition rate to 82%. As is illustrated in Fig. 13, the ability of the network to adapt to and overcome possible feature extraction errors in a single frame rises with the length of the tune in question, reaching more than 85% for tunes lasting more than a few seconds. Besides this, the network shows impressive ability to differentiate between half planes, especially in the case of activation (active vs. passive emotions).

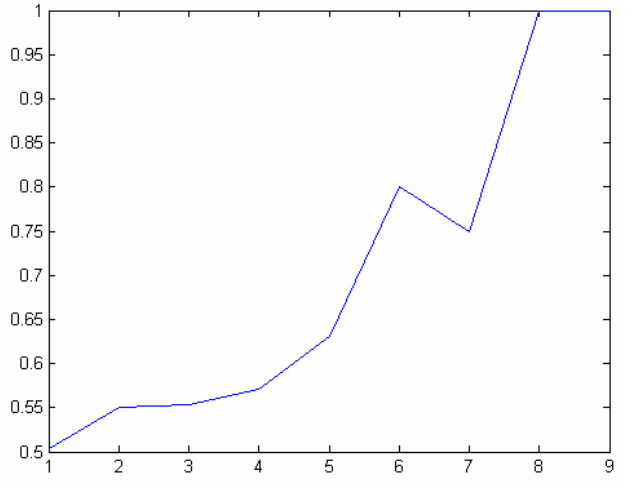


Fig. 13. Quadrant recognition rate over tune length (horizontal axis in intervals of 0.8 sec)

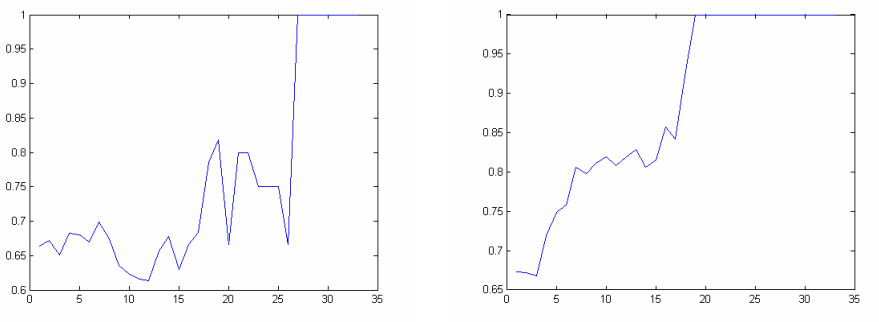


Fig. 14. (a) Positive-Negative recognition rate, (b) Active-Passive recognition rate, horizontal axis shows tune length in 0.2 sec units

7 Conclusions – Future Work

Naturalistic data goes beyond extreme emotions and concentrates on more natural emotional episodes that happen more frequently in everyday discourse. In this paper, we described a feature-based approach that tackles most of the intricacies of everyday audiovisual HCI and models the time-varying nature of these features in cases of expressivity. Most approaches focus on the detection of visual features in prerecorded, acted datasets and the utilization of machine learning algorithms to estimate the illustrated emotions. Even in cases of multimodality, features are fed into the machine learning algorithms without any real attempt to find structure and correlations between the features themselves and the estimated result. Neural networks are a nice solution to finding such relations, thus coming up with comprehensible connections between the input (features) and the output (emotion).

The fact that we use naturalistic and not acted data introduces a number of interesting issues, for example segmentation of the discourse in tunes. During the experiment, tunes containing a small number of frames (less than 5 frames, i.e. 0.2 seconds) were found to be error prone and classified close to chance level (not better than 37%). This is attributed to the fact that emotion in the speech channel needs at least half a second to be expressed via wording, as well as to the internal structure of the Elman network which works better with a short-term memory of ten frames. From a labeling point of view, ratings from four labelers are available; in some cases, experts would disagree in more than 40% of the frames in a single tune. In order to integrate this fact, the decision system has to take into account the interlabeller disagreement, by comparing this to the level of disagreement with the automatic estimation. One way to achieve this, is the modification of the Williams Index [13], which is used to this effect for the visual channel in [50].

A future direction regarding the features themselves is to model the correlation between phonemes and FAPs. In general, feature points from the mouth area do not contribute much when the subject is speaking; however, consistent phoneme detection could help differentiate expression-related deformation (e.g. a smile) to speech-related. Regarding the speech channel, the multitude of the currently detected features is hampering the training algorithms. To overcome this, we need to evaluate the importance/prominence of features so as to conclude on the influence they have on emotional transition.

Acknowledgments. This work has been funded by the FP6 Network of Excellence Humaine: Human-Machine Interaction Network on Emotion, URL:
http://www.emotion- research.net

References

1. A. Jaimes, Human-Centered Multimedia: Culture, Deployment, and Access, *IEEE Multimedia Magazine*, Vol. 13, No.1, 2006.
2. A. Mehrabian, Communication without Words, *Psychology Today*, vol. 2, no. 4, pp. 53-56, 1968.

3. A. Nogueiras, A. Moreno, A. Bonafonte and J.B. Mariño, Speech emotion recognition using hidden markov models. *Proceedings of Eurospeech*, Aalborg, Denmark, 2001.
4. A. Pentland, Socially Aware Computation and Communication, *Computer*, vol. 38, no. 3, pp. 33-40, 2005.
5. A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, Parameterized facial expression synthesis based on MPEG-4, *EURASIP Journal on Applied Signal Processing*, Vol. 2002, No 10, 2002, pp. 1021-1038.
6. A.J. Fridlund, *Human Facial Expression: An Evolutionary Perspective*, Academic Press, 1994.
7. B. Hartmann, M. Mancini, C. Pelachaud, Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis, In Computer Animation'02, Geneva, Switzerland. IEEE Computer Society Press, 2002.
8. C. Tomasi and T. Kanade, Detection and Tracking of Point Features, Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.
9. F. Freitag, E. Monte, Acoustic-phonetic decoding based on Elman predictive neural networks, *Proceedings of ICSLP 96*, Fourth International Conference on, Page(s): 522-525, vol.1.
10. FP5 IST ERMIS, Emotionally Rich Man-machine Intelligent System IST-2000-29319, http://www.image.ntua.gr/ermis
11. FP6 IST HUMAINE, Human-Machine Interaction Network on Emotion, 2004-2007, http://www.emotion-research.net
12. G. Caridakis, A. Raouzaiou, K. Karpouzis, S. Kollias. Synthesizing Gesture Expressivity Based on Real Sequences. Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 Conference, Genoa, Italy, 24-26 May.
13. G. W. Williams, Comparing the joint agreement of several raters with another rater", Biometrics, vol32, pp. 619-627, 1976.
14. H. G. Zimmermann, R. Grothmann, A. M. Schaefer, and Ch. Tietz. Identification and forecasting of large dynamical systems by dynamical consistent neural networks. In S. Haykin, T. Sejnowski J. Principe, and J. McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brain*. MIT Press, 2006.
15. I. Cohen, A. Garg, and T. S. Huang, Emotion Recognition using Multilevel-HMM, *NIPS Workshop on Affective Computing*, Colorado, Dec 2000.
16. I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T.S. Huang. Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. In Proc. Conf. on Computer Vision and Pattern Recognition, volume 1, pages 595–601, 2003.
17. J. Lien, Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity, Ph.D. dissertation, Carnegie Mellon University, Pittsburg, PA, 1998.
18. J. W. Young, Head and Face Anthropometry of Adult U.S. Civilians, *FAA Civil Aeromedical Institute*, 1963-1993 (final report 1993)
19. J. Weizenbaum, ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine, Communications of the ACM, Volume 9, Number 1, 1966, pp. 36-35.
20. J.L. Elman, Finding structure in time, *Cognitive Science*, vol. 14, 1990, pp. 179-211.
21. K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis and S. Kollias, Facial expression and gesture analysis for emotionally-rich man-machine interaction, in N. Sarris, M. Strintzis, (eds.), *3D Modeling and Animation: Synthesis and Analysis Techniques*, pp. 175-200, Idea Group Publ., 2004.

22. K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis and S. Kollias, Facial expression and gesture analysis for emotionally-rich man-machine interaction, N. Sarris, M. Strintzis, (eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques, pp. 175-200, Idea Group Publ., 2004.
23. K.M. Lam, H. Yan, Locating and Extracting the Eye in Human Face Images, *Pattern Recognition*, Vol.29, No.5, 1996, pp. 771-779.
24. K.R. Scherer, Adding the affective dimension: A new look in speech analysis and synthesis, In *Proc. International Conf. on Spoken Language Processing*, pp. 1808–1811, 1996.
25. L. Vincent, Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms, *IEEE Trans. Image Processing*, vol. 2, no. 2, 1993, pp. 176-201.
26. L.C. De Silva and P.C Ng, Bimodal emotion recognition, In *Proc. Automatic Face and Gesture Recognition*, pp. 332–335, 2000.
27. L.M. Wang, X.H. Shi, G.J. Chen, H.W. Ge, H.P. Lee, Y .C. Liang, Applications of PSO Algorithm and OIF Elman Neural Network to Assessment and Forecasting for Atmospheric Quality, ICANNGA 2005, 2005
28. L.S. Chen and T.S. Huang, Emotional expressions in audiovisual human computer interaction, In *Proc. International Conference on Multimedia and Expo*, pp. 423–426, 2000.
29. L.S. Chen, Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction, PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering, 2000.
30. M. H. Yang, D. Kriegman, N. Ahuja, Detecting Faces in Images: A Survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.24(1), 2002, pp. 34-58.
31. M. Pantic and L.J.M. Rothkrantz, Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
32. M. Pantic and L.J.M. Rothkrantz, Towards an Affect-sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.
33. M. Pantic, Face for Interface, in *The Encyclopedia of Multimedia Technology and Networking*, M. Pagani, Ed., Idea Group Reference, vol. 1, pp. 308-314, 2005.
34. M. Pantic, N. Sebe, J. Cohn, T. Huang, Affective Multimodal Human-Computer Interaction, *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 669 – 676, 2005.
35. Mathworks, Manual of Neural Network Toolbox for MATLAB
36. Murat Tekalp, Joern Ostermann, Face and 2-D mesh animation in MPEG-4, *Signal Processing: Image Communication* 15, Elsevier, pp. 387-421, 2000.
37. N. Sebe, I. Cohen, T.S. Huang, *Handbook of Pattern Recognition and Computer Vision*, World Scientific, January 2005
38. P. Ekman and W. Friesen, Pictures of Facial Affect, Palo Alto, CA: Consulting Psychologists Press, 1978.
39. P. Mertens, The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. in B. Bel & I. Marlien (eds.), *Proc. of Speech Prosody*, Japan, 2004.
40. P. Oudeyer, The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2):157-183, 2003.
41. R. Cowie and E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proc. International Conf. on Spoken Language Processing*, pp. 1989–1992, 1996.

42. R. Cowie, E. Douglas-Cowie, N.Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processign Magazine*, pp 33- 80, January 2001.
43. R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey and M. Schröder, FEELTRACE: An instrument for recording perceived emotion in real time, *ISCA Workshop on Speech and Emotion*, Northern Ireland, pp. 19-24, 2000.
44. R. Fransens, Jan De Prins, SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection, *Ninth IEEE International Conference on Computer Vision*, Volume 2, October 13 - 16, 2003.
45. R. W. Picard, *Affective Computing*, MIT Press, 1997.
46. R. W. Picard, Towards computers that recognize and respond to user emotion, *IBM Syst. Journal*, 39 (3–4), 705–719, 2000.
47. R.L. Hsu, M. Abdel-Mottaleb, Anil K. Jain, Face Detection in Color Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.24, No.5, May 2002
48. R.W. Picard, E. Vyzas, and J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.
49. S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, New York, 1994.
50. S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy network, *Neural Networks*, Elsevier, Vol. 18, Issue 4, May 2005, pp. 423-435
51. T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, S. Kollias, Emotion Analysis in Man-Machine Interaction Systems, Samy Bengio, Hervé Bourlard (Eds.), Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, Vol. 3361, 2004, pp. 318 - 328, Springer-Verlag.
52. U. Williams, K. N. Stevens, *Emotions and Speech: some acoustical correlates*, JASA 52, pp. 1238-1250, 1972.
53. Z. Zeng, J. Tu, M. Liu, T.S. Huang, Multi-stream Confidence Analysis for Audio-Visual Affect Recognition, *ACII 2005*, pp. 964-971.
54. Z. Zeng, Y. Hu, G. Roisman, Y. Fu, T. Huang, "Audio-visual, Emotion Recognition in Adult Attachment Interview, *this volume.*

Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning

Joost Broekens

Leiden Institute of Advanced Computer Science, Leiden University
Niels Bohrwerg 1, 2333CA Leiden, The Netherlands
broekens@liacs.nl

Abstract. Computer models can be used to investigate the role of emotion in learning. Here we present *EARL*, our framework for the systematic study of the relation between *e*motion, *a*daptation and *r*einforcement *l*earning (RL). EARL enables the study of, among other things, communicated affect as reinforcement to the robot; the focus of this chapter. In humans, emotions are crucial to learning. For example, a parent—observing a child—uses emotional expression to encourage or discourage specific behaviors. Emotional expression can therefore be a reinforcement signal to a child. We hypothesize that affective facial expressions facilitate robot learning, and compare a *social* setting with a *non-social* one to test this. The non-social setting consists of a simulated robot that learns to solve a typical RL task in a continuous grid-world environment. The social setting additionally consists of a human (parent) observing the simulated robot (child). The human's emotional expressions are analyzed in real time and converted to an additional reinforcement signal used by the robot; positive expressions result in reward, negative expressions in punishment. We quantitatively show that the "social robot" indeed learns to solve its task significantly faster than its "non-social sibling". We conclude that this presents strong evidence for the potential benefit of affective communication with humans in the reinforcement learning loop.

Keywords: Reinforcement Learning, Affect, Human-in-the-Loop.

1 Introduction

In humans, emotion influences thought and behavior in many ways [16][17][19][38]. For example, emotion influences how humans process information by controlling the broadness versus the narrowness of attention. Also, emotion functions as a social signal that communicates reinforcement of behavior in, e.g., parent-child relations. Computational modeling (including robot modeling) has proven to be a viable method of investigating the relation between emotion and learning [11][24], emotion and problem solving [3][6], emotion and social robots [7] (for review see [20]), and emotion, motivation and behavior selection [2][5][15][46]. Although many approaches exist and much work has been done on computational modeling of emotional influences on thought and behavior, none explicitly targets the study of the relation between emotion and learning using a complete end-to-end framework in a

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 113–132, 2007.

reinforcement learning context[1]. By this we mean a framework that enables systematic *quantitative* study of the relation between affect and RL in a large variety of ways, including (a) affect as reinforcement to the robot (both internally generated as well as socially communicated), (b) affect as perceptual feature to the robot (again internally generated and social), (c) affect resulting from reinforced robot behavior, and (d) affect as meta-parameters for the robot's learning mechanism. In this chapter we present such a framework. We call our framework *EARL*, short for the systematic study of the relation between *e*motion, *a*daptation and *r*einforcement *l*earning.

1.1 Affect as Reinforcement

In this chapter we specifically focus on the influence of socially communicated emotion on learning in a reinforcement learning context. This work is strongly related to research into interactive robot learning based on human advice or guidance [35][43]. We briefly review this area of research in Section 2. In the experimental part of this chapter we show, using our framework *EARL*, that human emotional expressions can be effectively used as additional reinforcement signal used by a simulated robot. Our experimental setting is as follows.

The robot's task is to optimize food-finding behavior while navigating through a continuous grid world environment. The grid world is not discrete, nor is an attempt made to define discrete states based on the continuous input. The gridworld contains walls, path and food patches. The robot perceives its direct surroundings as they are, and acts by turning and driving. We have developed an action-based learning mechanism that learns to predict values of actions based on the current perception of the agent (note that in this chapter we use the terms agent and robot interchangeably). Every action has its own Multi-Layer Perceptron (MLP) network (see also [28]) that learns to predict a modified version of the Q-value for that action [41]. The simulated robot does not use a separate training phase; we adopt the so-called certainty equivalence hypothesis [27].

We have used this setup to ensure that our simulation is as close as possible to a real world setting: continuous input directly fed into MLP networks. By doing so, we hope that observed robot behavior can be extrapolated to the real world: in theory, building the actual robot with appropriate sensors and actuators would suffice to replicate the results. We explain our modeling method in more detail in Section 4-6.

As mentioned above, we study the effect of a human's emotional expression on the learning behavior of the robot. As such the simulated robot uses the recognized emotion as a motivator for action, while the human uses its expression to signal the relevance of certain events (see also, [12]). In humans, emotions are crucial to learning. For example, a parent—observing a child—uses emotional expression to encourage or discourage specific behaviors. In this case, the emotional expression is used to setup an *affective communication channel* [36] and is used to communicate a reinforcement signal to a child. In this chapter we take *affect* to mean the positiveness versus the negativeness (*valence*) of a situation, object, etc. (see [11][38] and [39] for a more detailed argumentation of this point of view, and [47] for a detailed discussion

[1] Although the work by Gandanho [24] is a partial exception as it explicitly addresses emotion in the context of RL. However, this work does not address social human input and social robot output.

on the relation between valence and reinforcement learning). In our experiments, a human observes a simulated robot while the robot learns to find food. Affect in the human's facial expression is recognized by the robot in real time. As such, a smile is interpreted as communicating positive affect and therefore converted to a small additional reward (additional to the reinforcement the robot receives from its simulated environment). The expression of fear is interpreted as communicating negative affect and therefore converted to a small additional punishment. We call this the *social* setting. We vary between three types of social settings: one in which affect is a strong reinforcement but only for several learning trials; one in which affect is a moderate reinforcement for a longer period of time; and finally one in which affect is a moderate reinforcement while (in contrast to the first two types) the robot learns a social reward function that maps its perceived state to the social reinforcement. In the latter type, the robot can use its learned social reward function when the human stops giving social reinforcement. Finally, there is a *non-social* control setting to which the results of the social settings are compared. The non-social setting is a standard experimental reinforcement learning setup using the same elements as the social setups but without the social reinforcement.

We hypothesized that robot learning (in a RL context as described above) is facilitated by additional social reinforcement. Our experimental results support this hypothesis. We compared the learning performance of our simulated robot in the social and non-social settings, by analyzing averages of learning curves. The main contribution of this research is that it presents *quantitative evidence of the fact that a human-in-the-loop can boost learning performance in real-time by communicating reinforcement using facial expressions, in a non-trivial learning environment*. We belief this is an important result. It provides a solid base for further study of human mediated robot-learning in the context of real-world applicable reinforcement learning, using the communication protocol nature has provide for that purpose, i.e., emotional expression and recognition. As such, our results add weight to the view that robots can be trained and their behaviors optimized using *natural social cues*. This facilitates human-robot interaction and is relevant to human computing [32], to which we devote more attention in the discussion.

1.2 Chapter Layout

The rest of this chapter is structured as follows. In Section 2 we review related work. In Section 3 we discuss, in some detail, affect, emotion and how affect influences learning in humans. In Section 4 we briefly introduce *EARL*, our complete framework. In Section 5 we describe how communicated affect is linked to a social reinforcement signal. In Section 6, we explain our method of study (e.g., the grid-world, the learning mechanism). Section 7 discusses the results and Section 8 discusses these in a broader context and presents concluding remarks and future work.

2 Interactive Robot Learning

One of the main reasons for investigating natural ways of giving feedback to robots in order for them to be able to adapt themselves is *non-expert interaction*; the ability of

persons not familiar with machine learning to change the behavior of robots in a natural way [30][31][43]. Robots can learn from humans in a variety of ways, and humans want to teach robots in many different ways [43]. In general there are three types of robot teaching: *by example*, *by feedback* and *by guidance*. Our paper focuses on a method of interactive robot learning by feedback, but we briefly discuss all three approaches in this section.

2.1 Learning by Example

In the case of learning by example, robots learn behavior by imitating human behavior when that behavior is provided as an example (for review see [9]). The robot either imitates the behavior, or imitates getting towards the goal using the behavior as example. In the first case the behavior is leading, in the second the intention of the behavior is leading. Sometimes, robots can even learn to imitate. This is the case when the robot not only learns to imitate behavior, but also learns to imitate in the first place. The study presented in this chapter is not related to imitative behavior, as the human tutor in our study does not communicate to the robot examples of possible behaviors as to how to find the food (solve the task).

2.2 Learning by Feedback

Our study falls into the category *learning by feedback*. In this case the robot learns by receiving performance feedback from the human in the form of an additional reinforcement signal [10][26][30][31][34[45]. Such signals can come in many forms. For example in the study by Isbell et al. [26], their social chatter bot *Cobot* learns the information preferences of its chat partners, by analyzing the chat messages for explicit and implicit reward signals (e.g., positive or negative words). These signals are then used to adapt its model of providing information to that chat partner. So, Cobot effectively uses social feedback as reward, as does our simulated robot. However, there are several important differences. Cobot does not address the issue of a human observer parenting the robot using affective communication. Instead, it learns based on reinforcement extracted from words used by the user during the chat sessions in which Cobot is participating. Also, Cobot is not a real-time behaving robot, but a chat robot. As a consequence, time constraints related to the exact moment of administering reward or punishment are less important. Finally, Cobot is restricted regarding its action-taking initiative, while our robot is continuously acting, with the observer reacting in real-time.

Thrun et al. [31] describe how their museum tour-guide robot *MINERVA* learns how to attract attention from humans as well as how to optimize tours (in terms of the time available and the exhibits visited). An important task this robot has to learn is how to attract people to its tours. The reward signal for learning this task is based on the amount of people being close to the robot; too close, however, represents a negative reinforcement, so the robot should adapt its attention-attracting behavior to maximize having a reasonably sized group waiting around it. In this study, a non-intrusive measure has been used as basis for the reinforcement signal. This is comparable with the approach used by Mitsunaga et al. [30]. They also use non-intrusive signals (robot-human distance, gaze aversion, and body repositioning) as

reinforcement signal to adapt the robot-human distance to a comfortable one. Obviously, two key differences between these studies and ours exist: we explicitly use the affective channel (facial expression) to communicate reinforcement, and we analyzed whether this reinforcement signal helps the simulated robot to solve a task that is not related to human-robot interaction per se.

Studies that are particularly related to ours are the ones by Papudesi and Huber [34][35]. They investigate if a composite reward function (composed of the normal reinforcement given by the environment and reinforcement based on human advice) enhances robot learning of a navigation problem in a grid-based maze. The human-based part of the reward function is done in quite a clever way. The robot is given a set of advice instructions on where to go first or what choice to make at junctions. This advice is in terms of state-action pairs, so, a certain action in a certain state (representing a location in the maze) is given a slight selection bias. All biases together form a bias function (over the state-action space) that can be translated to a user-based reinforcement function. This user-based reinforcement is the first part of the reward function, and is added to the environment-based reinforcement. Together this forms the composite reward function. This composite function is used for training. The interesting part is that by using this two-step approach of administering user reward, formal analysis of the maximum permissible user-reward values is possible. The authors have shown boundaries for the human advice such that several problems related to additional user rewards can be overcome, problems such as "looping" (due to intermediate user reinforcement, the robot keeps looping through the same state-action pairs). The key difference between their approach and ours is that we use the facial expression to communicate the reward function, that we have a continuous state representation (see Section 6) and that we administer the user's reinforcement directly, without a bias function. It would be interesting to merge both approaches and use facial expression to feed the bias function as defined in [34].

2.3 Learning by Guidance

Learning by guidance is a relatively new approach to interactive human-robot learning [42][43][45]. Guidance can be differentiated from feedback and imitation (example) in the following way. While feedback gives intentional information after the fact, guidance gives intentional information before the fact (*anticipatory reinforcement*; [44]). For example, smiling at a robot after it has taken the right turn towards the food (our study), is quite different from proposing a certain turn to the robot before it has chosen itself [44].

While imitation assumes a sequence of behaviors that lead towards a goal state, guidance is about future-directed learning cues and as such is much broader defined. For example, showing how to tie shoelaces is very different from drawing a child's attention to the two edges when stuck in the beginning. In general, robot guidance is about directing attention, communicating motivational intentions, and proposing actions [43].

For example in the work by Thomaz and Breazeal [43], the authors show an interesting way in which guidance can be added to a standard reinforcement learning mechanism. A human can advice the agent to pay more attention to a specific object in the problem environment (in this case a learning-to-cook environment, called

Sophie's kitchen, with a simulated robot). The guidance is transformed into an action-selection bias, such that the simulated robot selects actions that have to do with the advised object with higher probability. As such, this approach resembles the one by Papudesi and Huber [34]: the behavioral bias is given at the level of state-action pairs, not directly at the level of reward and punishment. The main reason why biasing action-selection as done by [43] can best be seen as guidance and biasing action-states as done by [34] can best be seen as feedback, is the way both studies use the human advice. In the former (T&B), the advice is immediately integrated into the action-selection, and guides the robot's next actions, while in the latter (P&H) the advice is first translated to an additional reward for certain state-action pairs and the resulting composite reward function is used for training the robot.

In a real sense, guidance by biasing action-selection can be seen as narrowing-down attention towards certain objects or features. By biasing action-selection, certain actions have a higher change of being selected than others. This kind of human-robot interaction can help solve exploration-exploitation issues in very large state spaces, as the human can guide de robot into a useful direction, while the robot still has the opportunity to explore [44].

3 Affect Influences Learning

In this chapter we specifically focus on the influence of socially communicated affect on learning, i.e., on affectively communicated feedback. Affect and emotion are concepts that lack a single concise definition, instead there are many [37]. Therefore we first explain our meaning to these concepts.

3.1 Emotion and Affect

In general, the term emotion refers to a set of—in social animals—naturally occurring phenomena including facial expression, motivation, emotional actions such as fight or flight behavior, a tendency to act, and—at least in humans—feelings and cognitive appraisal (see, e.g., [40]). An emotional state is the combined activation of instances of a subset of these phenomena, e.g., angry involves a tendency to fight, a typical facial expression, a typical negative feeling, etc. Time is another important aspect in this context. A short term (intense, object directed) emotional state is often called an *emotion*; while a longer term (less intense, non-object directed) emotional state is referred to as *mood*. The direction of the emotional state, either positive or negative, is referred to as *affect* (e.g., [39]). Affect is often differentiated into two orthogonal (independent) variables: *valence*, a.k.a. pleasure, and *arousal* [19][39]. Valence refers to the positive versus negative aspect of an emotional state. Arousal refers to the activity of the organism during that state, i.e., physical readiness. For example, a car that passes you in a dangerous manner on the freeway, immediately (*time*) elicits a strongly negative and highly arousing (*affect*) emotional state that includes the expression of anger and fear, feelings of anger and fear, and intense cognitive appraisal about what could have gone wrong. On the contrary, learning that one has missed the opportunity to meet an old friend involves cognitive appraisal that can negatively influence (*affect*) a person's mood for a whole day (*time*), even though the

associated emotion is not necessarily arousing (*affect*). Eating a piece of pie is a more positive and biochemical example. This is a bodily, emotion-eliciting event resulting in mid-term moderately-positive affect. Eating pie can make a person happy by, e.g., triggering fatty-substance and sugar-receptor cells in the mouth. The resulting positive feeling typically is not of particularly strong intensity and certainly does not involve particularly high or low arousal, but might last for several hours.

3.2 Emotional Influences

Emotion influences thought and behavior in many ways. For example, at the neurological level, malfunction of certain brain areas not only destroys or diminishes the capacity to have (or express) certain emotions, but also has a similar effect on the capacity to make sound decisions [17] as well as on the capacity to learn new behavior [4]. Behavioral evidence suggests that the ability to have sensations of pleasure and pain is strongly connected to basic mechanisms of learning and decision-making [4]. These findings indicate that brain areas important for emotions are also important for "classical" cognition and instrumental learning.

At the level of cognition, a person's belief about something is updated according to the associated emotion: the current emotion is used as information about the perceived object [14][21], and emotion is used to make the belief resistant to change [22]. Ergo, emotions are "at the heart of what beliefs are about" [23].

Emotion plays a role in the regulation of the amount of information processing. For instance, Scherer [40] argues that emotion is related to the continuous checking of the environment for important stimuli. More resources are allocated to further evaluate the implications of an event, only if the stimulus appears important enough. Furthermore, in the work of Forgas [21] the relation between emotion and information processing strategy is made explicit: the influence of mood on thinking depends on the strategy used. In addition to this, it has been found that positive moods favor creative thoughts as well as integrative information processing, while negative moods favor systematic analysis of incoming stimuli (e.g. [1][25]).

Emotion also regulates behavior of others. Obvious in human development, expression (and subsequent recognition) of emotion is important to communicate (dis)approval of the actions of others. This is typically important in parent-child relations. Parents use emotional expression to guide behavior of infants. Emotional interaction is essential for learning. Striking examples are children with an autistic spectrum disorder, typically characterized by a restricted repertoire of behaviors and interests, as well as social and communicative impairments such as difficulty in joint attention, difficulty recognizing and expressing emotion, and lacking of a social smile (for review see [13]). Apparently, children suffering from this disorder have both a difficulty in building up a large set of complex behaviors *and* a difficulty understanding emotional expressions and giving the correct social responses to these. This disorder provides a clear example of the interplay between learning behaviors and being able to process emotional cues.

As argued by Breazeal and Brooks [8], human emotion is crucial to understanding others, as well as ourselves, and this could very well be two equally crucial functions for robot emotions.

3.3 Socially Communicated Affect

To summarize, emotion and mood influence thought and behavior in a variety of ways, e.g., a person's mood influences processing style and attention, emotions influence how one thinks about objects, situations and persons, and emotion is related to learning new behaviors.

In this study we focus on the role of affect in guiding learning in a social human-robot setting. We use affect to denote the positiveness versus negativeness of a situation. We ignore the arousal a certain situation might bring. As such, positive affect characterizes a situation as good, while negative affect characterizes that situation as bad (e.g., [39]). Further, we use affect to refer to the *short term* timescale: i.e., to emotion. We hypothesize that affect communicated by a human observer can enhance robot learning. In our study we assume that the recognition of affect translates into a reinforcement signal. As such, the robot uses a *social reinforcement* in addition to the reinforcement it receives from its environment while it is building a model of the environment using reinforcement learning mechanisms. In the following sections we first explain our framework after which we detail our method and discuss results and further work.

4 EARL: A Computational Framework to Study the Relation Between Emotion, Adaptation and Reinforcement Learning

To study the relation between emotion, adaptation and reinforcement learning, we have developed an end-to-end framework. The framework consists of four parts:

- An emotion recognition module, recognizing emotional facial expression in real time.
- A reinforcement learning agent to which the recognized emotion can be fed as input.
- An artificial emotion module slot; this slot can be used to plug into the learning agent different models of emotion that produce the artificial emotion of the agent as output. The modules can use all of the information that is available to the agent (such as action repertoire, reward history, etc.). This emotion can be used by the agent as intrinsic reward, as metalearning parameter, or as input for the expression module.
- An expression module, consisting of a robot head with the following degrees of freedom: eyes moving up and down, ears moving up and down on the outside, lips moving up and down, eyelids moving up and down on the outside, and RGB eye colors.

Emotion recognition is based on quite a crude mechanism based upon the face tracking abilities of OpenCV [48]. Our mechanism uses 9 points on the face, each defined by a blue sticker: 1 on the tip of the nose, 2 above each eyebrow, 1 at each mouth corner and 1 on the upper and lower lip. The recognition module is configured to store multiple prototype point constellations. The user is prompted to express a certain emotion and press space while doing so. For every emotional expression (in the case of our experiment neutral, happy and afraid), the module records the

positions of the 9 points relative to the nose. This is a prototype point vector. After configuration, to determine the current emotional expression in real time the module calculates a weighted distance from the current point vector (read in real-time from a web-cam mounted on the computer screen) to the prototype vectors. Different points get different weights. This results in an error measure for every prototype expression. This error measure is the basis for a normalized vector of recognized emotion intensities. The recognition module sends this vector to the agent (e.g., neutral 0.3, happy 0.6, fear 0.1). Our choice of weights and features has been inspired by work of others (for review see [33]). Of course the state of the art in emotion recognition is more advanced than our current approach. However, as our focus is affective learning and not the recognition process per se, we contented ourselves with a low fidelity solution (working almost perfectly for neutral, happy and afraid, when the user keeps the head in about the same position).

Note that we do not aim at generically recognizing detailed emotional expressions. Instead, we tune the recognition module to the individual observer to accommodate his/her personal and natural facial expressions. The detail with which this is done reflects our experimental needs: extract positive and negative reward signals from the observer's face. In a real-world scenario with observers and robots autonomously acting next to each other, a more sophisticated mechanism is needed to correctly read reward signals from the observers. Such a mechanism needs to be multi-modal.

The reinforcement learning agent receives this recognized emotion and can use this in multiple ways: as reinforcement, as information (additional state input), as metaparameter (e.g., to control learning rate), and as social input directly into its emotion model. In this chapter we focus on social reinforcement, and as such focus on the recognized emotion being used as additional reward or punishment. The agent, its learning mechanism and how it uses the recognized emotion as reinforcement are detailed in Sections 5 and 6.

The artificial emotion model slot enables us to plug in different emotion models based on different theories to study their behavior in the context of reinforcement learning. For example, we have developed a model based on the theory by Rolls [38], who argues that many emotions can be related to reward and punishment and the lack thereof. This model enables us to see if the agent's situation results in a plausible (e.g., scored by a set of human observers) emotion emerging from the model. By scoring the plausibility of the resulting emotion, we can learn about the compatibility of, e.g., Rolls' emotion theory with reinforcement learning. However, in the current study we have not used this module, as we focus on affective input as social reinforcement.

The emotion expression part is a physical robot head. The head can express an arbitrary emotion by mapping it to its facial features, again according to a certain theory. Currently our head expresses emotions according to the Pleasure Arousal Dominance (PAD) model by Mehrabian [29]. We have a continuous mapping from the 3-dimensional PAD space to the features of the robot face. As such we do not need to explicitly work with emotional categories or intensities of the categories. The mapping appears to work quite well, but is in need of validation (again using human observers). We have not used the robot head for the studies reported upon in this chapter.

We now describe in detail how we coupled the recognized human emotion to the social reinforcement signal for the robot. Then we explain in detail our adapted reinforcement learning mechanism (such that it enabled learning in continuous environments), and our method of study as well as our results.

5 Emotional Expressions as Reinforcement Signal

As mentioned earlier, emotional expressions and facial expressions in particular can be used as social cues for the desirability of a certain action. In other words, an emotional expression can express reward and punishment if directed at an individual. We focus on communicated affect, i.e., the positiveness versus negativeness of the expression. If the human expresses a smile (happy face) this is interpreted as positive affect. If the human expresses fear, this is interpreted as negative affect. We interpret a neutral face as affectless.

We have studied the mechanism of communicated affective feedback in a human-robot interaction setup. The human's face is analyzed (as explained above) and a vector of emotional expression intensities is fed to the learning agent. The agent takes the expression with the highest intensity as dominant, and equates this with a *social reinforcement* of, e.g., 2 (happy), −2 (fear) and 0 (neutral). It is important to realize that this is a simplified setup, as the human face communicates much more subtle affective messages and at the very least is able to communicate the degree of reward and punishment. For example, fear and anger are two distinct negative emotions that have different meaning and different action-tendencies. Fear involves a tendency to avoid and is not directed at an individual (although it can be caused by an individual), while anger involves a tendency to approach and is outwardly directed at someone else. A little bit of anger might be interpreted as a little bit of punishment, while a lot of anger better be interpreted as "don't ever do this again". However, to investigate our hypothesis (affective human feedback increases robot learning performance) the just described mechanism is sufficient. For the sake of simplicity, in this experiment we take fear as a "prototype for negative affective facial expression" and happiness as a "prototype for positive affective facial expression".

The social reinforcement, called r_{social}, is simply added to the "normal" reinforcement the agent receives from its environment (together forming a composite reinforcement). So, if the agent walks on a path somewhere in the gridworld, it receives a reinforcement (say 0), but when the user smiles, the resulting actual reinforcement becomes 2, while if the user looks afraid, the resulting reinforcement becomes −2.

6 Method

To study the impact of social reinforcement on robot learning, we have used our framework in a simulated continuous gridworld. In this section we explain our experimental setup.

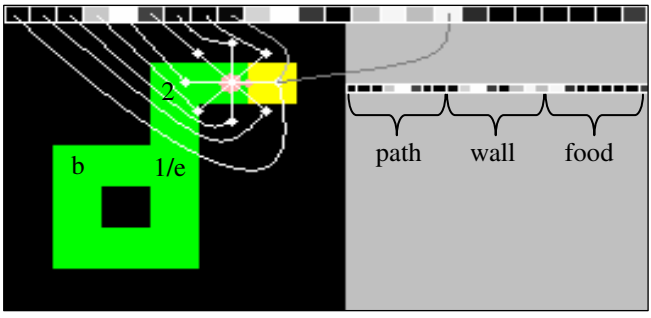


Fig. 1. The experimental gridworld. The agent is the "circle with nose" in the top right of the maze, where the nose denotes its direction. The 8 white dots denote the points perceived by the agent. These points are connected to the elements of state *s* (neural input to the MLPs used by the agent) as depicted. This is repeated for all possible features, in our case: path (gray), wall (black), and food (light gray), in that order. The *e* denotes the cell in which social reinforcement can be administered through smiling or expression of fear, the *1* and *2* denote key locations at which the agent has to learn to differentiate its behavior, i.e., either turn left (*1*) or right (*2*). The agent starts at *b*. The task enforces a non-reactive best solution (by which we mean that there is no direct mapping from reinforcement to action that enables the agent to find the shortest path to the food). If the agent would learn that turning right is good, it would keep walking in circles. If the agent learns that turning left is good, it would not get to the food.

6.1 Continuous Gridworld as Test Environment

A simulated robot (agent) "lives" in a continuous gridworld environment consisting of wall, food and path patches (Figure 1). These are the features of the world observable by the agent. The agent cannot walk on walls, but can walk on path and food. Walls and path are neutral (have a reinforcement of 0.0), while food has a reinforcement of 10. One cell in the grid is assumed to be a 20 by 20 object. Even though wall, path and food are placed on a grid, the world is continuous in the following sense: the agent has real-valued coordinates, moves by turning or walking in a certain direction using an arbitrary speed (in our experiments set at 3), and perceives its direct surroundings (within a radius of 20) according to its looking direction (one out of 16 possible directions). The agent uses a "relative eight neighbor metric" meaning that it perceives features of the world at 8 points around it, with each point at a distance of 20 from the center point of the agent and each point at an interval of 1/4 PI radians, with the first point always being exactly in front of it (Figure 1). The state perceived by the agent (its percept) is a real-valued vector of inputs between 0 and 1; each input is defined by the relative contribution of a certain feature in the agent-relative direction corresponding to the input. For example, if the agent sees a wall just in front of it (i.e., the center point of a wall object is exactly at a distance of 20 as measured from the current agent location in its looking direction) the first value in its perceived state would be equal to 1. This value can be anywhere between 0 and 1 depending on the distance of that point to the feature. For the three types of features, the agent thus has 3x8=24 real-valued inputs between 0 and 1 as its perceived world state *s* (Figure 1). As such the agent can approach objects (e.g., a wall) from a large number of possible angles and positions, with every intermediate position being possible. For

all practical purposes, the learning environment can be considered continuous. We did not define discrete states based on the continuous input to facilitate learning. Instead we chose to use the perceived state as is, to maximize potential transferability of our experimental results to real-world robot learning.

6.2 Reinforcement Learning in Continuous Environments

Reinforcement learning in continuous environments introduces several important problems for standard RL techniques, such as Q learning, mainly because a large number of potentially similar states exist as well as a very long path length between start and goal states making value propagation difficult. We now briefly explain our adapted RL mechanism. As RL in continuous environments is not specifically the topic of the chapter we have left out some of the rational for our choices.

The agent learns to find the path to the food, and optimizes this path. At every step the agent takes, the agent updates its model of the expected benefit of a certain action as follows. It learns to predict the value of actions in a certain perceived state s, using an adapted form of Q learning. The value function, $Q_a(s)$, is approximated using a Multi-Layer Perceptron (MLP), with 3x8=24 input, 24 hidden, and one output neuron(s), with s being the real-valued input to the MLP, a the action to which the network belongs, and the output neuron converging to $Q_a(s)$. As such, every action of the agent (5 in total: forward, left, right, left and forward, right and forward) has its own network. The output of the action networks are used as action values in a standard Boltzmann action-selection function [41]. An action network is trained on the Q value—i.e., $Q_a(s) \leftarrow Q_a(s)+\alpha(r+\gamma Q(s')-Q_a(s))$ —where r is the reward resulting from action a in state s, s' is the resulting next state, $Q(s')$ the value of state s', α is the learning rate and γ the discount factor [41]. The learning rate equals 1 in our experiments (because the learning rate of the MLP is used to control speed of learning, not α), and the discount factor equals 0.99. To cope with a continuous gridworld, we adapted standard Q learning in the following way:

First, the value $Q_a(s)$ used to train the MLP network for action a is topped such that $min(r, Q_a(s')) <= Q_a(s) <= max(r, Q(s'))$. As a result, individual $Q_a(s)$ values can never be larger or smaller than any of the rewards encountered in the world. This enables a discount factor close to or equal to 1, needed to efficiently propagate back the food's reward through a long sequence of steps. In continuous, cyclic, worlds, training the MLP on normal Q values using a discount factor close to 1 can result in several problems not further discussed here.

Second, per step of the agent, we train the action-state networks not only on $Q_a(s) \leftarrow Q_a(s)+\alpha(r+\gamma Q(s')-Q_a(s))$ but also on $Q_a(s') \leftarrow Q_a(s')$. The latter seems unnecessary but is quite important. RL assumes that values are propagated *back*, but MLPs generalize while trained. As a result, training an MLP on $Q_a(s)$ also influences its value prediction for s' in the same direction, just because the inputs are very close. In effect, part of the value is actually propagated *forward*; credit is partly assigned to what comes next. This violates the RL assumption just mentioned. Note that the value $Q(s')$ is predicted using another MLP, called the value network, that is trained in the same way as the action networks using the topped-off value and forward propagation compensation.

Third, for the agent to better discriminate between situations that are perceptually similar, such as position "1" and "2" in Figure 1, for each action-network the agent also uses a second network trained on the value of *not* taking the action. This network is trained when other actions are taken but not when the action to which the "negation" network belongs is taken. In effect, the agent has two MLPs per action. This enables the agent to better learn that, e.g., "right" is good in situation "2" but *not* in situation "1". Without this "negation" network, the agent learns much less efficient (results not shown). To summarize, our agent has 5 actions, it has 11 MLPs in total: one to train $Q(s)$, 5 to train $Q_a(s)$ and 5 to train $-Q_a(s)$. All networks use forward propagation compensation and a topped-off value to train upon. The MLP predictions for $Q_a(s)$ and $-Q_a(s)$ are simply added, and the result is used for action-selection.

6.3 Social vs. Non-social Learning

To study the effect of communicated affect as social reinforcement, we created the following setup. First an agent is trained without social reinforcement. The agent repeatedly tries to find the food for 200 trials, i.e., one *run*. The agent continuously learns and acts during these trials. To facilitate learning, we use a common method to vary the MLP learning rate and the Boltzmann action selection β derived from simulated annealing. The Boltzmann β equals to $3+(trial/200)*(6-3)$, effectively varying from 3 in the first trial to 6 in the last. The MLP learning rate equals to $0.1-(trial/200)*(0.1-0.001)$ effectively varying from 0.1 in the first trial to 0.001 in the last. We repeated the experiment 200 times, resulting in 200 runs. Average learning curves are plotted for these 200 runs using a linear smoothing factor equal to 6 (Figure 2).

Second, a new agent is trained *with* social reinforcement, i.e., a human observer looking at the agent with his/her face analyzed by the agent, translating a smile to a social reward and a fearful expression to a social punishment. Again, average learning curves are plotted using a linear smoothing factor equal to 6, but now based on the average per trial over 15 runs (Figure 2). We experimented with three different social settings: (a) a moderate social reinforcement, r_{human}, from trial 20 to 30, where the social reinforcement is either −0.5 or 0.5 (happy vs. fearful, respectively); (b) a strong social reinforcement, r_{human}, from trial 20 to 25 where social reinforcement is either −2 or 2, i.e., more extreme social reinforcement but for a shorter period; (c) a social reinforcement, r_{human}, from trial 29 to 45 where social reinforcement is either −2 or 2 while (in addition to settings *a* and *b*) the agent trains an additional MLP to predict the direct social reinforcement, r_{human}, based on the current state s. The MLP is trained to learn $R_{social}(s)$ as given by the human reinforcement r_{human}. After trial 45, the direct social reinforcement from the observer, r_{human}, is replaced by the learned social reinforcement $R_{social}(s)$. So, during the critical period (the trial intervals mentioned) of social setting *a*, *b* and *c*, the total reinforcement is a composite reward equal to $R(s)+r_{human}$. Only in setting *c*, and only after the critical period until the end of the run, the composite reward equals $R(s)+R_{social}(s)$. In all other periods, the reinforcement is as usual, i.e., $R(s)$. As a result, in setting *c* the agent can continue using an additional social reinforcement signal that has been learned based on what its human tutor thinks about certain situations.

The process of giving affective feedback to a reinforcement learning agent appeared to be quite a long, intensive and attention absorbing experience. As a result, it was physically impossible to observe the agent during all trials in the entire gridworld (after 2 hours of smiling to a computer screen one is exhausted *and* has burning eyes and painful facial muscles). To be able to test our hypothesis, we restricted direct social input to (I) a critical learning period defined in terms of a start and end trail (as discussed above), and (II) the cell indicated by *e* (Figure 1). Only when the agent moves around in this cell *and* is in a social input trial, the simulation speed of the experiment is set to one action per second enabling affective feedback.

7 Results

The results clearly show that learning is facilitated by social reinforcement. In all three social settings (Figure 2a, b and c) the agent needs fewer steps to find the food during the trials in which the observer provides assistance to the agent by expressing positive or negative affect. Interestingly, at the moment the observer stops reinforcing, the agent gradually looses the learning benefit it had accumulated. This is independent of the size of the social reinforcement (both social learning curves in Figure 2a and b show dips that eventually return to the non-social learning curve).

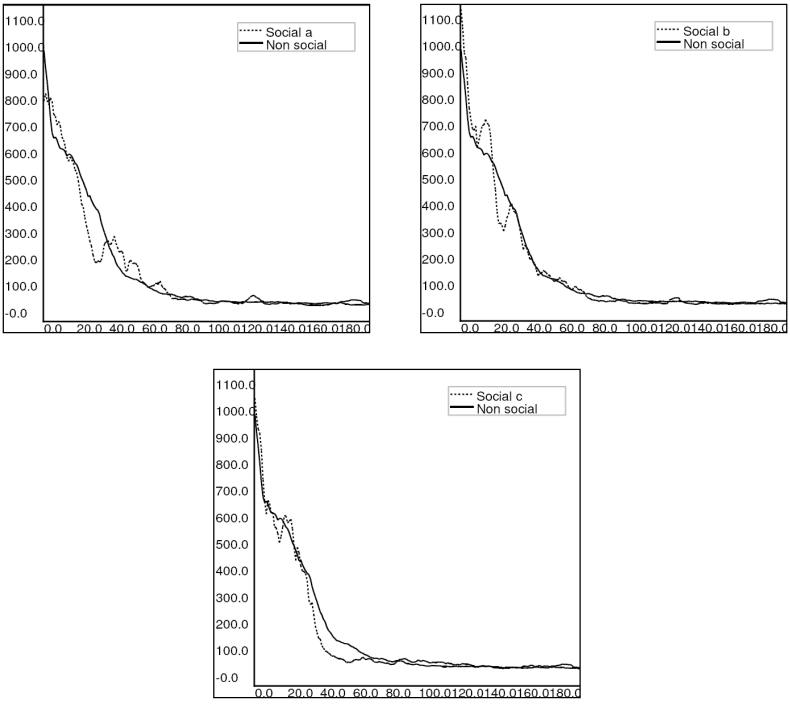


Fig. 2. Results of the learning experiments. From top to bottom showing the difference between the non-social setting and social setting *a*, *b*, and *c* respectively.

Loss of learning benefit as observed in social settings *a* and *b* (learning curve moving up, starting at trial 30 and 25 respectively) can be easily explained. The social reinforcement was not given long enough for the agent to internalize the path to the food (i.e., propagate back the food's reward to the beginning of the path). As soon as the observer stops reinforcing, the agent starts to forget these rewards, i.e., the MLPs are again trained to predict values as they are without social reinforcement. So, either the observer should continue to reinforce until the agent has internalized the solution, or the agent needs to be able to build a representation of the social reward function and use it when direct social reinforcement is not available. We have experimented with the second (social setting *c*): we enabled the agent to learn the social reward function. Now the agent uses direct social reinforcement at the emotional input spot (*e*, Figure 1) during the critical period, and uses its social reward prediction, $R_{social}(s)$, when direct social reinforcement stops. Results clearly show that the agent is now able to keep the benefit it had accumulated from using social reinforcement (Figure 2c). These results show that a combination of using social reinforcement and learning a social reward function facilitates robot learning, by enabling the robot to quicker learn the optimal solution to the food due to the direct social reinforcement as well as keep that solution by using its learned social reward function when social reinforcement stops.

8 Conclusion, Discussion and Future Work

Our results show that affective interaction in human-in-the-loop learning can provide a significant benefit to the efficiency of a reinforcement learning robot in a continuous grid world. We believe our results are particularly important to human-robot interaction for the following reasons. First, advanced robots such as robot companions, robot workers, etc., will need to be able to adapt their behavior according to human feedback. For humans it is important to be able to give such feedback in a natural way, e.g., using emotional expression. Second, humans will not want to give feedback all the time, it is therefore important to be able to define critical learning periods as well as have an efficient social reward system. We have shown the feasibility of both. Social input during the critical learning periods was enough to show a learning benefit, and the relatively easy step of adding an MLP to learn the social reward function enabled the robot to use the social reward when the observer is away.

We have specifically used an experimental setup that is compatible with a real-world robot: we have used continuous inputs and MLP-based training of which it is known that it can cope with noise and generalize over training examples. As such we believe our results can be generalized to real-world robotics. However, this most certainly needs to be experimented with.

A related issue is that the training time needed to learn our (arguably) simple task is quite long. This is also due to the representational format of the environment resulting in long state-action sequences to the goal state with states that resemble each other quite a lot. A discrete world with less, more discriminative, states can use a standard form of reinforcement learning and will show a more marked effect of intermediate social reinforcement.

Future work includes a broader evaluation of the EARL framework including its ability to express emotions generated by an emotional model plugged into the RL agent. Further, we envision to experiment with controlling metaparameters (such as exploration/exploitation and learning rate) based on the agent's internal emotional state or social rewards [3][11][18]. Currently we use simulated annealing-like mechanisms to control these parameters. Further, the agent could try to learn what an emotional expression predicts. In this case, the agent would use the emotional expression of the human in a more pure form (e.g., as a real-valued vector of facial feature intensities as part of its perceived state *s*). This might enable the agent to learn what the emotional expression means for itself instead of simply using it as reward.

As mentioned earlier, no distinction is made between different facial expressions that portray positive emotions or negative emotions. For example, no difference is made between the meaning of sadness versus anger. Thus, the current setup is highly simplified regarding the type of information that can be communicated through the affective channel. Future work includes a coupling between other reinforcement learning parameters and other aspects of facial expressions. For example, fear portrays a future danger and as such could be used by the agent to reconsider its current actions used for action-selection. Anger communicates a form of blame: the agent should have known better. This could be used to reevaluate (and perhaps internally simulate) a stored sequence of recent interactions in order to come up with an alterative, more positive, outcome than the current one.

Another way to extend this work is proposed by Thomaz, Hoffman and Breazeal [45]. Currently, the simulated robot is influenced by the human observer only at a certain spot in the maze. This is quite limited. However, it has been proposed [7][45] that human tutors could very well use a robot's behavioral cues as a signal to intervene with the learning process. For example, agent-to-teacher signals such as gaze, gesture and hesitation could be used by the tutor to, for example, propose actions to the agent, give motivational feedback etc. [45]. In our setup, we have often observed behavior that can be characterized as hesitation (e.g., the simulated robot switching between turning left and right but not deciding on really making the complete turn to take a branch in the maze). It would be interesting to allow the human tutor to influence the robot more freely, and to investigate if (1) humans tend to recognize hesitation behavior in our setup and (2) if affective feedback can still be used in these circumstances or whether a more guidance-based approach is need at these hesitation moments.

With regards to human computing [32], our work shows two things: *real-time* natural feedback (in our case, facial expressions) is feasible and desirable for robot learning, and, a personalized reward function can be learned based on this real-time interaction. This is relevant to two issues in human computing: *dynamics* and *learning/education*. Our work quantitatively shows that dynamic interaction with a simulated learning robot, using natural means of input (face) instead of traditional means (keyboard, mouse) enhances learning of robot behavior. Our interpretation of *dynamics* in this paper is somewhat different from the *dynamics* as meant in [32], i.e., the dynamics of the behavioral cue itself and the problem of deciphering these dynamics. Nevertheless, dynamic interaction is an important issue to human computing: one would not want to have to stop the robot before feedback can be communicated. Regarding the second issue, i.e., how to learn the user specific

meaning to an interactive pattern, we have shown that it is feasible to learn in real-time a personalized social reward function that the robot can use to train itself. A straightforward addition to this work would be to learn multiple social reward functions depending on, e.g., user and task context, such that the robot can select which reward function to use in what context. This would help the robot adapt to new contexts in a lazy and unsupervised way [32]. As we have already mentioned earlier, one could use a strategy in which the robot does not directly couple reward to facial expressions, but instead *learns* to couple facial features to reward. Now, a robot could first learn what different expressions mean to different users, and subsequently use the appropriate reward function to adapt its behavior.

Finally, a somewhat futuristic possibility is actually quite close: affective Robot-Robot interaction. Using our setup, it is quite easy to train one robot in a certain environment (parent), make it observe an untrained robot in that same environment (child), and enable it to express its emotion as generated by its emotion model using its robot head, an expression recognized and translated into social rewards by the child robot. Apart from the fact that it is somewhat dubious if such a setup is actually useful (why not send the social reward as a value through a wireless connection to the child), it would enable robots to use the same communication protocol as humans.

Regarding the "usefulness" argument just put forward, it seems to apply to our experiment as well. Why didn't we just simulate affective feedback by pushing a button for positive reward and pushing another for negative reward (or even worse, by simulating a button press)? From the point of view of the robot this is entirely true, however, from the point of view of the human—and therefore the point of view of the human-robot interaction—not at all. Humans naturally communicate social signals using there face, not by pushing buttons. The process of expressing an emotion is quite different from the process of pushing a button, even if it was only for the fact that it takes more time and effort to initiate the expression and that the perception of an expression is the perception of a process not a discrete event (like a button press). In a real-world scenario with a mobile robot in front of you it would be quite awkward to have to push buttons instead of just smile when you are happy about its behavior. Further it would be quite useful if the robot could recognize you being happy or sad and gradually learn to adapt its behavior even when you did not intentionally give it a reward or punishment. Abstracting away from the actual affective interaction patterns between the human and the robot in our experiment would have rendered the experiment almost completely trivial. Nobody would be surprised to see that the robot learns better if an intermediate reward is given halfway its route towards food. Our aim was to investigate if affective communication can enhance learning in a reinforcement learning setting. Taking out the affective part would have been quite strange indeed.

Acknowledgments. We would like to sincerely thank Pascal Haazebroek and all the students who helped us develop the *EARL* system, thereby making this research possible. Joris Slob, Chris Detweiler, Sylvain Vriens, Koen de Geringel, Hugo Scheepens, Remco Waal, Arthur de Vries, Pieter Jordaan, Michiel Helvensteijn, Rogier Heijligers, Willem van Vliet, you were great!

References

1. Ashby, F. G., Isen, A. M., Turken, U.: A Neuro-psychological theory of positive affect and its influence on cognition. Psychological Review 106 (3) (1999) 529-550
2. Avila-Garcia, O., Cañamero, L.: Using hormonal feedback to modulate action selection in a competitive scenario. In: From Animals to Animats 8: Proc. 8th Intl. Conf. on Simulation of Adaptive Behavior. MIT Press, Cambridge MA (2004) 243-252
3. Belavkin, R. V.: On relation between emotion and entropy. In: Proc. of the AISB'04 Symposium on Emotion, Cognition and Affective Computing. AISB Press (2004) 1-8
4. Berridge, K. C.: Pleasures of the brain. Brain and Cognition 52 (2003) 106-128
5. Blanchard, A. J., Cañamero, L.: Modulation of exploratory behavior for adaptation to the context. In: Proc. of the AISB'06 Symposium on Biologically Inspired Robotics (Biro-net). AISB Press (2006) 131-137
6. Botelho, L. M., Coelho, H.: Information processing, motivation and decision making. In: Proc. 4th International Workshop on Artificial Intelligence in Economics and Management. (1998)
7. Breazeal, C.: Affective interaction between humans and robots. In: J. Keleman, P. Sosik (eds): Proc. of the ECAL 2001. LNAI, Vol. 2159. Springer-Verlag, Berlin Heidelberg New York (2001) 582-591
8. Breazeal, C., Brooks R.: Robot emotion: A functional perspective. In: J.-M. Fellous, M. Arbib (eds.): Who needs emotions: The brain meets the robot. Oxford University Press USA (2004) 271-310
9. Breazeal, C., Scassellati, B.: Robots that imitate humans. Trends in Cognitive Sciences 6(11) (2002) 481-487
10. Breazeal, C., Velasquez, J.: Toward teaching a robot `infant' using emotive communication acts. In: Edmonds, B., Dautenhahn, K. (eds.): Socially Situated Intelligence: a workshop held at SAB'98, Zürich. University of Zürich Technical Report (1998) 25-40
11. Broekens, J., Kosters, W. A., Verbeek, F. J.: On emotion, anticipation and adaptation: Investigating the potential of affect-controlled selection of anticipatory simulation in artificial adaptive agents. In press (2007)
12. Cañamero, D.: Designing emotions for activity selection. Dept. of Computer Science Technical Report DAIMI PB 545. University of Aarhus Denmark (2000)
13. Charman, T., Baird, G.: Practitioner review: Diagnosis of autism spectrum disorder in 2- and 3-year-old children. Journal of Child Psychology and Psychiatry 43(3) (2002) 289-305
14. Clore, G. L. Gasper, K.: Feeling is believing: Some affective influences on belief. In: Frijda, N., Man-stead A. S. R., Bem, S. (eds.): Emotions and Beliefs. Cambridge Univ. Press, Cambridge UK (2000) 10-44
15. Cos-Aguilera, I., Cañamero, L., Hayes, G. M., Gillies, A.: Ecological integration of affordances and drives for behaviour selection. In: Proc. of the Workshop on Modeling Natural Action Selection. AISB Press (2005) 225-228
16. Custers, R., Aarts, H.: Positive affect as implicit motivator: On the nonconscious operation of behavioral goals. Journal of Personality and Social Psychology 89(2) (2005) 129-142
17. Damasio, A. R.: Descartes' error. Penguin Putnam, New York NY (1994)
18. Doya, K.: Metalearning and neuromodulation. Neural Networks 15 (4) (2002) 495-506
19. Dreisbach, G., Goschke, K.: How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. Journal of Experimental Psychology: Learning, Memory, and Cognition 30(2) (2004) 343-353

20. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robots and Autonomous Systems 42 (2003) 143-166
21. Forgas, J. P.: Feeling is believing? The role of processing strategies in mediating affective influences in beliefs. In: Frijda, N., Manstead A. S. R., Bem, S. (eds.): Emotions and Beliefs. Cambridge University Press, Cambridge UK (2000) 108-143
22. Frijda, N. H., Mesquita, B.: Beliefs through Emotions. In: Frijda, N., Manstead A. S. R., Bem, S. (eds.): Emotions and Beliefs. Cambridge University Press, Cambridge UK (2000) 45-77
23. Frijda, N. H., Manstead, A. S. R., Bem. S.: The influence of emotions on beliefs. In: Frijda, N., Man-stead A. S. R., Bem, S. (eds.): Emotions and Beliefs. Cambridge University Press, Cambridge UK (2000) 1-9
24. Gandanho, S. C.: Learning behavior-selection by emotions and cognition in a multi-goal robot task. Journal of Machine Learning Research 4 (2003) 385-412
25. Gasper, K., Clore, L. G.: Attending to the big picture: Mood and global versus local processing of visual information. Psychological Science 13(1) (2002) 34-40
26. Isbell, C. L. Jr., Shelton, C. R., Kearns, M., Singh, S., Stone, P.: A social reinforcement learning agent. In: Proceedings of the fifth international conference on Autonomous agents. ACM (2001) 377-384
27. Kaelbling, L. P., Littman, M. L., Moore, A. W.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research 4 (1996) 237-285
28. Lin, L. J.: Reinforcement learning for robots using neural networks. Doctoral dissertation. Carnegie Mellon University, Pittsburgh (1993)
29. Mehrabian, A.: Basic Dimensions for a General Psychological Theory. OG&H Publishers, Cambridge Massachusetts (1980)
30. Mitsunaga, N., Smith, C., Kanda, T., Ishiguro, H., Hagita, N.: Robot behavior adaptation for human-robot interaction based on policy gradient reinforcement learning. In: Proc. Of the International Conference on Intelligent Robots and Systems (IROS). IEEE Press (2005) 218-225
31. Thrun, S., Bennewitz, M., Burgard, W., Cremers, A. B., Dellaert, F., Fox, D., Hähnel, D., Rosenberg, C. R., Roy, N., Schulte, J., Schulz, D.: A tour-guide robot that learns. In: Burgard, W., Christaller, T., Cremers, A.B. (eds.): Proc. of the 23rd Annual German Conference on Artificial Intelligence: Advances in Artificial Intelligence. LNAI, Vol. 1701. Springer-Verlag, London UK (1999) 14-26
32. Pantic, M., Pentland, A., Nijholt, A., Huang, T. S.: Human computing and machine understanding of human behavior: A Survey. Proc. ACM Int'l Conf. Multimodal Interfaces (2006) 239-248
33. Pantic, M., Rothkranz, L. J. M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1424-1445
34. Papudesi, V. N., Huber, M.: Learning from reinforcement and advice using composite reward functions. In: Proc. Of the 16th International FLAIR Conference. AAAI (2003) 361-365
35. Papudesi, V. N., Huber, M.: Interactive refinement of control policies for autonomous robots. In: Proc. Of the 10th IASTED International Conference on Robotics and Applications, Honolulu HI. IASTED (2004)
36. Picard, R. W.: Affective Computing. MIT Press, Cambridge MA (1997)
37. Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C. Cavallo, D., Machover, T., Resnick, M., Roy, D., Strohecker, C.: Affective learning — A manifesto. BT Technology Journal 22(4) (2004) 253-269

38. Rolls, E. T.: Précis of The brain and emotion. Behavioral and Brain Sciences 23 (2000) 177-191
39. Russell, J. A.: Core affect and the psychological construction of emotion. Psychological Review 110(1) (2003) 145-72
40. Scherer, K. R.: Appraisal considered as a process of multilevel sequential checking. In: K. R. Scherer, A. Schorr, T. Johnstone (eds.): Appraisal processes in emotion: Theory, Methods, Research. Oxford Univ. Press, New York NY (2001) 92-120
41. Sutton, R., Barto, A.: Reinforcement learning: An introduction. MIT Press, Cambridge MA (1998)
42. Ogata, T., Sugano, S., Tani, J.: Open-end human robot interaction from the dynamical systems perspective: Mutual adaptation and incremental learning. In: Orchard, R., Yang, C., Ali, M. (eds.): 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. LNCS, Vol. 3029. Springer (2004) 435-444
43. Thomaz, A.L., Breazeal, C.: Teachable characters: User studies, design principles, and learning performance. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.): Proc. of the 6th International Conference on Intelligent Virtual Agents (IVA 2006). LNCS, Vol. 4133, Springer (2006) 395-406
44. Thomaz, A. L., Breazeal, C.: Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In: Proc. of the 21st National Conference on Artificial Intelligence. AAAI Press (2006b)
45. Thomaz, A.L., Hoffman, G., Breazeal, C.: Real-time interactive reinforcement learning for robots. In: Proc. of AAAI Workshop on Human Comprehensible Machine Learning. Pittsburgh, PA (2005)
46. Velasquez, J. D.: A computational framework for emotion-based control. In: SAB'98 Workshop on Grounding Emotions in Adaptive Systems (1998)
47. Wright, I.: Reinforcement learning and animat emotions. In: From Animals to Animats 4: Proc. of the 4th International Conference on the Simulation of Adaptive Behavior (SAB): MIT Press, Cambridge MA (1996) 272-284
48. OpenCV: http://www.intel.com/technology/computing/opencv/index.htm

Trajectory-Based Representation of Human Actions

Antonios Oikonomopoulos[1], Ioannis Patras[2], Maja Pantic[1], and Nikos Paragios[3]

[1]Imperial College London, 180 Queensgate, SW7 2AZ London, UK
{aoikonom,m.pantic}@imperial.ac.uk
[2]Department of Electronic Engineering, Queen Mary University, London, UK
I.Patras@elec.qmul.ac.uk
[3]Ecole Centrale de Paris, Grande Voie des Vignes, 92 295 Chatenay-Malabry, France
nikos.paragios@ecp.fr

Abstract. This work addresses the problem of human action recognition by introducing a representation of a human action as a collection of short trajectories that are extracted in areas of the scene with significant amount of visual activity. The trajectories are extracted by an auxiliary particle filtering tracking scheme that is initialized at points that are considered salient both in space and time. The spatiotemporal salient points are detected by measuring the variations in the information content of pixel neighborhoods in space and time. We implement an online background estimation algorithm in order to deal with inadequate localization of the salient points on the moving parts in the scene, and to improve the overall performance of the particle filter tracking scheme. We use a variant of the Longest Common Subsequence algorithm (LCSS) in order to compare different sets of trajectories corresponding to different actions. We use Relevance Vector Machines (RVM) in order to address the classification problem. We propose new kernels for use by the RVM, which are specifically tailored to the proposed representation of short trajectories. The basis of these kernels is the modified LCSS distance of the previous step. We present results on real image sequences from a small database depicting people performing 12 aerobic exercises.

1 Introduction

With an ever-increasing role of computers and other digital devices in our society, one of the main foci of the research in Artificial Intelligence should be on Emerging Human-Machine Systems. A related, crucial issue is that of Human-Machine Interaction (HMI). A long-term goal in HMI research is to approach the naturalness of human-human interaction. This means integrating 'natural' means that humans employ to interact with each other into the HMI designs. With this motivation, automatic speech recognition and synthesis have been the topics of research for decades. Recently, other human interactive modalities such as facial and body gestures have also gained interest as potential modes of HMI. The analysis of what is present in a scene is an essential issue in the

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 133–154, 2007.

development of natural Human-Machine interfaces. Who is in the scene, what is (s)he doing, and how is (s)he doing it, are essential questions that should be answered if natural interfaces that are non-obtrusive and informative for the user are to be realized. Vision offers the means to achieve this. It also represents an essential link to the creation of systems which are able to adapt to the affective state of their user, leading in this way to an affect-sensitive interaction between the user and the machine. Particularly for ambient intelligence, anticipatory interfaces, and human computing, the key is the ease of use - the ability to unobtrusively sense certain behavioral cues of the users and to adapt automatically to their typical behavioral patterns and the context in which they act [1]. Sensing and interpretation of human behavioral cues play an important role and are extremely relevant for the development of applications in the fields of security (video surveillance and monitoring), natural multimodal interfaces, augmented reality, smart rooms, object-based video compression and driver assistance. Tremendous amount of work has been done in the field in recent years [2],[3].

In this work we present an unsupervised method for the representation of the activity taking place in a scene. This method is based on the detection of salient points in space and time, that correspond to regions with a significant amount of activity. Subsequently, we track these points in time using a state-estimation approach in order to reach a representation based on short trajectories. We test the proposed method using real image sequences of subjects performing several aerobic exercises. It can be used, however, to represent any type of activity, including hand gestures, gait, extraction of motion patterns etc. Possible applications lie in the area of e-health, where the development of non-stationary, non-intrusive, non-invasive monitoring inside and outside the clinical environment is essential, due to demanding patients, aging population and rising costs. The method can be realized as an adaptive system that will be able to monitor and assess the correctness of the performed exercise, and will provide an appropriate alternative (senior) fitness plan, assisting in this way nurses,physical therapists and family members. The system can also be configured for use at home, to accommodate elderly but otherwise healthy patients or patients suffering from conditions like rheumatism and chronic pain.

2 Related Work

2.1 Tracking

The main objective of tracking is to estimate the state x_k (e.g. position, pose) given all the measurements $z_{1:k}$ up to the current time instant k. In a probabilistic framework, this translates in the construction of the a posteriori probability $p(x_k|z_{1:k})$. Theoretically, the optimal solution in case of Gaussian noise in the measurements is given by the Kalman filter [4], which yields the posterior being also Gaussian. Kalman filters and their variants, like the Extended (EKF) and the Unscented Kalman Filters (UKF) [5], [6], [7] have been extensively used for a

variety of tracking applications [8], [9]. However, in nonlinear and non-Gaussian state estimation problems Kalman filters can be significantly off.

To overcome the limitations of Kalman filtering, the classical particle filtering algorithm, or so-called Condensation algorithm was proposed [10], [11]. The main idea behind particle filtering is to maintain a set of possible solutions called particles. Each particle is associated with a weight, the latter expressing the likelihood of the particle being the actual solution. By maintaining a set of solutions instead of a single estimate as is done by Kalman filtering, particle filters are more robust to missing and inaccurate data. The major drawback of the classic Condensation algorithm, however, is that a large amount of particles might be wasted because they are propagated into areas with small likelihood. In order to overcome this problem, a number of variants to the original algorithm have been proposed, having as a common characteristic the goal of achieving a more optimal allocation of new particles. Since particle weights determine how the particles are being resampled, the likelihood function has an essential influence on the tracking performance [12]. Several attempts have been made in order to adjust the way new particles are assigned, through the use of kernels [13], [14], [15], orientation histograms [16] or special transformations like Mean Shift [17].

Despite the improvement in the tracking performance of the previous methods, the inherent problem of the classic condensation algorithm, that is, the propagation of particles in areas of small likelihood is not sufficiently addressed. In order to effectively deal with this issue, the Auxiliary Particle Filtering (APF) algorithm was proposed by Pitt and Shephard [18]. The APF algorithm operates in two steps. At first, particles are propagated and their likelihood is evaluated. Subsequently, the algorithm chooses again and propagates the particles according to the likelihood of the previous step. Since the introduction of the APF algorithm, a number of variants have been proposed in order to address different issues. In [19] a modified APF tracking scheme is proposed for the tracking of deformable facial features, like mouth and eye corners. The method uses an invariant color distance that incorporates a shape deformation term as an observation model to deal with the deformations of the face. In order to take into account spatial constraints between tracked points, the particle filter with factorized likelihoods is proposed in [20], where the spatial constraints between different facial features are pre-learned and the proposed scheme tracks constellations of points instead of a single point, by taking into account these constraints.

Particle filters are often used within a template tracking framework. The object's appearance is captured in the first frame of an image sequence and subsequently tracked throughout the end of the sequence. The underlying assumption behind template tracking is that the object will not significantly change its appearance throughout the duration of the video. This assumption, however, is not realistic, since an object can undergo several rotations, deformations or partial occlusions, making the template no longer an accurate model of the appearance of the object. A simple but rather naive solution to this problem is to update the template at every frame with a new template corresponding to the tracked

position of the object. This approach, however, leads to error accumulation, as small errors are constantly introduced in the location of the template. As a result, the template eventually drifts away from the object and in the most cases gets stuck on the static background of the scene. A compromising solution between these two extremes is to partially update the template, as the weighted average (e.g. 90-10 %) of the current and the initial template, a process often called exponential forgetting. Although this solution offers a somewhat more robust tracking, by allowing the template to adapt, it does not avoid error accumulation, and there is still a high probability that the template will eventually drift away from the object.

Matthews *et al* specifically address the drift problem in [21]. The tracked template is updated at every frame, while maintaining the initial template specified in the first frame. To eliminate drift, the new template is aligned every time to the initial one using a gradient descent rule. This strategy, however, is most suitable for tracking rigid objects (e.g. cars). For objects whose appearance changes over time, the authors adopt an approach of template tracking with Active Appearance Models (AAM). The appearance model and the template are updated at every time instance, leading to more robust tracking algorithm. A similar framework is presented in [22], where a set of adaptive appearance models are used for motion-based tracking. The appearance model used consists of three components. The stable component (S) is used to capture the behavior of temporally stable and slowly varying image observations, the data outlier or 'lost' component (L) is used to capture data outliers due to failures in tracking, occlusion or noise and finally the 'wandering' component (W) is used to model sudden changes in the appearance of the object. The parameters of the model are adjusted online via EM and the system is tested in tracking scenarios where a high degree of partial object occlusion occurs. Finally, in [23] a Support Vector Machine (SVM) is used in order to provide an initial guess for an object position in the first frame. The position of the initial guess in subsequently refined so that a local maximum of the SVM score is achieved. The whole framework is called Support Vector Tracking (SVT) and is implemented in moving vehicle tracking scenarios.

2.2 Human Activity Tracking and Recognition

A major component in human computing research is localization and tracking of the human body, either as a whole or as a part (e.g. head,limbs). Especially for the purposes of scene analysis and activity recognition, body tracking has received a lot of attention in the last few years. Due to its high degree of freedom (usually 28-60), body tracking is inherently a very difficult problem. Because of that, it calls upon sophisticated tracking algorithms, that can address the problem of high dimensionality. Furthermore, large appearance changes, occlusion between body parts, and the absence of typical appearance due to clothing, pose additional problems that need to be dealt with.

In contrast to rigid objects, tracking of articulated objects is inherently a much more difficult problem, mainly due to the high number of degrees of freedom

that are involved. Accurate human body tracking, in particular, is an extremely important aspect for human computing applications. A possible strategy for estimating the configuration of articulated objects is sequential search, in which a number of parameters are initially estimated and, assuming that this estimation is correct, the values of several other parameters are determined. For instance, Gavrila and Davis in [24] first locate the torso of the human body and then use this information in order to initialize a search for the limbs. This approach, however, only works for specific views and is very sensitive to self-occlusion that is, occlusion between different body parts. A similar approach is presented in [25], where a particle filtering framework is used for the purposes of hand tracking. For the same purpose, Cipolla *et al* [26] propose a view-based hierarchical probabilistic tracking framework that can deal with changes in view and self occlusions. The system uses edge and color cues in order to estimate the likelihood function of the hand position and configuration and subsequently a Bayesian filtering framework that performs the tracking. In [27] a particle filtering approach is adopted for articulated hand tracking. The tracker is guided by attractors, precollected training samples of possible hand configurations whose observations are known, while the whole process is modeled by a Dynamic Bayesian Network. A Bayesian Network is also adopted in [28] in order to model the existing constraints between the different parts of the human body. These constraints are learned using Gaussian Mixture Models (GMM) and training is done using motion-capture frames of walking data as the ground truth. Observations are based on multi-scale edge and ridge filters while the whole process is assisted with a pooled background model derived by the set of training images. In [29] a Dynamic Markov Network is utilized instead to model the relations between body parts and tracking is done using an sequential Monte Carlo algorithm. A similar approach is presented in [30], where an elastic model is used to represent relations and constraints between the limbs and a Nonparametric Belief Propagation (NBP) algorithm for the purpose of tracking. In [31] a combination of particle filters and Hidden Markov Models (HMM) is used for tracking and recognition respectively, of articulated hand gestures. Appearance-based models are learned for the non-rigid motion of the hand and a filtering method is used for the underlying rigid motion. Both treatments are unified into a single Bayesian framework. A similar approach is implemented in [32], where arm gestures are recognized as a sequence of body poses. The latter are recognized via edge matching and HMMs are used in order to extract the gestures from the pose sequences. HMMs are also used in [33] for recognizing pointing gestures. Skin information is used to localize the hands and the head of the subject in a scene and a multiple hypothesis scheme is used for the tracking. Subsequently, an HMM-based approach is adopted for recognizing the gestures.

Articulated object tracking, and particularly human body tracking suffer from dimensionality issues, an inherent problem whenever there is a large number of degrees of freedom. This fact makes the use of tracking algorithms like particle filters rather impractical. The reason for this is that a very large number of particles is required in order to represent the posterior function in a sufficient way, making

this kind of tracking algorithms slow and computationally expensive. The problem becomes even more prominent whenever real-time performance is required, such as in monitoring applications, virtual trainers or augmented reality applications. In order to deal with this issue, a number of different techniques have been proposed, either by constraining the configuration space [24] or by restricting the range of the movements of the subject [34]. These approaches, however, greatly reduce the generality of the implemented trackers, making them impractical in real applications. Eigenspace decomposition [35] and principal component analysis [36] offer an interesting alternative for dimensionality reduction. In [37], a modified particle filtering approach is used in order to reduce the complexity of human body tracking. The main characteristic of the utilized tracker is its ability to avoid local maxima in the tracking by incorporating a search based on simulated annealing, and thus called annealed particle filter. Apart from dimensionality reduction techniques, several researchers have attempted to modify the way classical tracking algorithms work in order to achieve computational efficiency and real-time tracking performance. A simple example are the earlier mentioned kernel-based particle filters [13], [14], [15], [38] or particle filters that use special transformations, as in [16], [17]. These methods attempt to limit the number of required particles for efficient tracking, effectively reducing the computational complexity of their algorithms. Finally, an interesting approach for real-time tracking and recognition of hand actions is presented in [39],[40]. The motion of the hand is extracted using skin cues and is subsequently tracked using the Mean-Shift Tracking scheme of [38]. The spatiotemporal curvatures of the extracted trajectories are used in order to represent the actions performed. The local maxima of these curvatures are view-invariant and are used for image sequence alignment and matching of the actions.

2.3 Unsupervised Representation and Recognition of Actions

Despite their extreme usefulness, tracking methods consist of only a fraction of the methods used for capturing the activity going on in a scene. While trackers mainly concentrate on tracking the state of an object at any time instant, a variety of other methods have been proposed that deal the problem in a more abstract or unsupervised manner [41], [42]. An interesting work is presented in [43], where human actions are treated as three-dimensional shapes in the space-time volume. The method utilizes properties of the solution to the Poisson equation to extract space-time features of the moving human body, such as local space-time saliency, action dynamics, shape structure and orientation. Subsequently, spectral clustering is used in order to group similar actions. In [44], long video sequences are segmented in the time domain by detecting single events in them. The detection is completely unsupervised, since it is done without any prior knowledge of the types of events, their models, or their temporal extent. The method can be used for event-based indexing even when only one short example-clip is available. Unsupervised methods for learning human motion patterns are also presented in [45], [46]. In these methods, the human body is modeled as a

triangulated graph. The model is learned in an unsupervised manner from unlabelled data using global and local features, either dynamic or appearance-based. The authors effectively handle occlusion by modeling the missing parts as hidden variables. The parameters of the assumed models are being estimated using EM. Finally, a Hidden Markov Model approach for action recognition is presented in [47]. The activity in a scene is represented by codewords called movelets. Each movelet is a collection of the shape, motion and occlusion of image patches corresponding to the main parts of the body. Recognition is done using HMMs, by estimating the most likely sequence of codewords and the action that took place in a sequence.

2.4 Overview of the Proposed Method

A wide variety of activity recognition methods use edge and color cues [16], [14] or some form of markers [48],[49] in order to assist initialization and the overall operation of tracking or recognition processes. In order to avoid the use of markers, an interesting alternative could be the use of interesting points for tracker initialization. According to Haralick and Shapiro [50] an interesting point is a) distinguishable from its neighbors and b) its position is invariant with respect to the expected geometric transformation and to radiometric distortions. Gilles introduces the notion of saliency in terms of local signal complexity or unpredictability in [51]. Kadir and Brady [52] extend the original Gilles algorithm and estimate the information content of pixels in circular neighborhoods at different scales in terms of the entropy. Local extremes of changes in the entropy across scales are detected and the saliency of each point at a certain scale is defined in terms of the entropy and its rate of change at the scale in question.

In this work, we propose a human action recognition algorithm that is based on the detection and tracking of spatiotemporal features in given image sequences. We do this by extending in the temporal direction the salient feature detector developed in [52]. The detected salient points correspond to peaks in activity variation such as the edges of a moving object. Similar to [43], we treat the action as three-dimensional events, by detecting the salient points in the space-time domain. Contrary to [14], [37], [28] and [29] that use models to represent the human body, we propose an entirely unsupervised method based on the detected salient features in order to represent the moving parts of the body. In this sense, the concept of our method resembles the one in [44], where detection is done without prior knowledge of the types of events, their models, or their temporal extent. Like in [52], we automatically detect the scales at which the entropy achieves local maxima and form spatiotemporal salient regions by clustering spatiotemporal points with similar location and scale. We derive a suitable distance measure between sets of salient regions, which is based on the Chamfer distance, and we optimize this measure with respect to a number of temporal and scaling parameters. In this way we achieve invariance against scaling and we eliminate the temporal differences between the representations. We extend our previous work on salient points presented at [53] by using the

detected salient regions in order to initialize a tracking scheme based on the auxiliary particle filter, proposed in [18]. Each image sequence is then represented as a set of short trajectories. The spatiotemporal coordinates of the points that consist the extracted trajectories are appropriately transformed according to the parameters that were estimated in the Chamfer distance optimization step. We use the adaptive background estimation algorithm presented in [54] in order to model the background in the available sequences and to improve the overall quality of the implemented tracking scheme. We use a variant of the Longest Common Subsequence algorithm (LCSS) that was proposed in [55],[56] in order to compare different sets of trajectories. We use Relevance Vector Machines [57] in order to address the classification problem. We propose new kernels for use by the RVM, which are specifically tailored to the proposed short trajectory representation. The basis of these kernels is the modified LCSS distance of the previous step. The novelty of the proposed method lies in the unsupervised nature of representation of the actions. Since we don't use any model, the method can be easily extended and used for a variety of different actions, ranging from full-body actions to single gestures.

The remainder of the paper is organized as follows: In section 3, the spatiotemporal feature detector used is described, along with the proposed space-time warping technique. In section 4, the auxiliary particle filter that was used is briefly analyzed along with the background estimation model that was utilized. In section 5 the proposed kernel-based recognition method is described. In section 6, we present our experimental results, and in section 7, final conclusions are drawn.

3 Spatiotemporal Salient Points

3.1 Spatiotemporal Saliency

Let us denote by $N_c(s, \boldsymbol{v})$ the set of pixels in an image I that belong to a circular neighborhood of radius s, centered at pixel $\boldsymbol{v} = (x, y)$. In [52], in order to detect salient points in static images, Kadir and Brady define a saliency measure $y_D(s, \boldsymbol{v})$ based on measuring changes in the information content of N_c for a set of different circular radiuses (i.e. scales). In order to detect spatiotemporal salient points at peaks of activity variation we extend the Kadir's detector by considering cylindrical spatiotemporal neighborhoods at different spatial radiuses s and temporal depths d. More specifically, let us denote by $N_{cl}(\boldsymbol{s}, \boldsymbol{v})$ the set of pixels in a cylindrical neighborhood of scale $\boldsymbol{s} = (s, d)$ centered at the spatiotemporal point $\boldsymbol{v} = (x, y, t)$ in the given image sequence. At each point $\boldsymbol{v}$ and for each scale $\boldsymbol{s}$ we will define the spatiotemporal saliency $y_D(\boldsymbol{s}, \boldsymbol{v})$ by measuring the changes in the information content within $N_{cl}(\boldsymbol{s}, \boldsymbol{v})$. Since we are interested in activity within an image sequence, we consider as input signal the convolution of the intensity information with a first-order Gaussian derivative filter. Formally, given an image sequence $I_0(x, y, t)$ and a filter G_t, the input signal that we use is defined as:

$$I(x, y, t) = G_t * I_0(x, y, t)\,. \tag{1}$$

For each point $\boldsymbol{v}$ in the image sequence, we calculate the Shannon entropy of the signal histogram in a cylindrical neighborhood $N_s(\boldsymbol{s}, \boldsymbol{v})$ around it. That is,

$$H_D(s, d, \boldsymbol{v}) = -\sum_{q \in D} p(q, s, d, \boldsymbol{v}) \log p(q, s, d, \boldsymbol{v}) \,, \tag{2}$$

The set of scales at which the entropy is peaked is given by:

$$\begin{aligned} \hat{S}_p = \{(s, d) : H_D(s-1, d, \boldsymbol{v}) < H_D(s, d, \boldsymbol{v}) > H_D(s+1, d, \boldsymbol{v}) \\ \wedge H_D(s, d-1, \boldsymbol{v}) < H_D(s, d, \boldsymbol{v}) > H_D(s, d+1, \boldsymbol{v})\} \end{aligned} \tag{3}$$

The saliency measure at the candidate scales is given by:

$$y_D(s, d, \boldsymbol{v}) = H_D(s, d, \boldsymbol{v}) W_D(s, d, \boldsymbol{v}), \quad \forall\, (s, d) \in \hat{S}_p \,, \tag{4}$$

The first term of eq. 4 is a measure of the variation in the information content of the signal. The weighting function $W_D(s, \boldsymbol{v})$ is a measure of how prominent the local maximum is at s, and is given by:

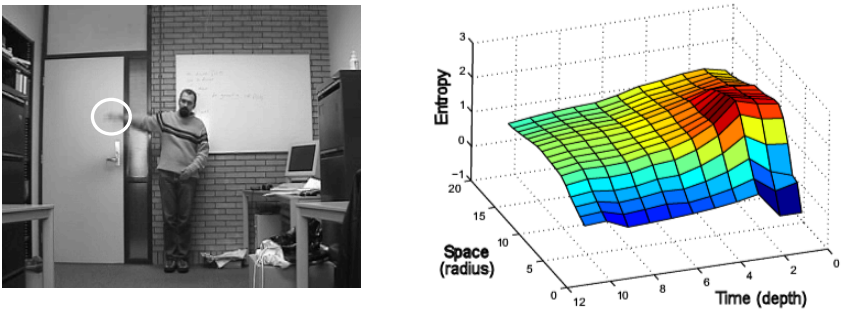
$$\begin{aligned} W_D(s, d, \boldsymbol{v}) = \frac{s^2}{2s-1} \sum_{q \in D} |p(q, s, d, \boldsymbol{v}) - p(q, s-1, d, \boldsymbol{v})| \\ + d \sum_{q \in D} |p(q, s, d, \boldsymbol{v}) - p(q, s, d-1, \boldsymbol{v})|, \forall\, (s, d) \in \hat{S}_p \,, \end{aligned} \tag{5}$$

where the values in front of each summation in the right part of eq. 5 are normalization factors. When a peak in the entropy for a specific scale is distinct, then the corresponding pixel probability density functions at the neighboring scales will differ substantially, giving a large value to the summations of eq. 5 and thus, to the corresponding weight value assigned. On the contrary, when the peak is smoother, then the summations in eq. 5 will have a smaller value. Let us note that we considered cylindrical neighborhoods for simplicity reasons. However, more complicated shapes, such as elliptical neighborhoods at different orientations and with different axes ratios could be considered.

In Fig. 1(a), a single frame from a sample image sequence is presented, where the subject is raising its right hand. By selecting as origin the center pixel of the drawn white circle, we apply a number of cylindrical neighborhoods of various scales in the sequence and we calculate the corresponding entropy values. The result is shown in Fig. 1(b), where the various entropy values are plotted with respect to the radiuses and depths of the corresponding cylindrical neighborhoods. The scale which corresponds to the distinct peak of the plot is considered candidate salient scale, and is assigned a saliency value, according to eq. 4.

3.2 Salient Regions

The analysis of the previous section leads to a set of candidate spatiotemporal salient points $S = \{(\boldsymbol{s}_i, \boldsymbol{v}_i, y_{D,i})\}$, where $\boldsymbol{v}_i = (x, y, t)$, $\boldsymbol{s}_i$ and $y_{D,i}$ are respectively, the position vector, the scale and the saliency value of the feature point with index i. In order to achieve robustness against noise we follow a similar approach as that in [52] and develop a clustering algorithm, which we apply to



(a) (b)

Fig. 1. (a) Single frame from a sample image sequence where the subject is raising its right hand and (b) the corresponding entropy plot as a function of the spatial radius and temporal depth of all the applied cylindrical neighborhoods. The origin of all the applied cylindrical neighborhoods is the center of the white circle in (a).

the detected salient points. By this we define salient regions instead of salient points, the location of which should be more stable than the individual salient points, since noise is unlikely to affect all of the points within the region in the same way. The proposed algorithm removes salient points with low saliency and creates clusters that are a) well localized in space, time and scale, b) sufficiently salient and c) sufficiently distant from each other. The steps of the proposed algorithm can be summarized as follows:

1. Derive a new set S_T from S by applying a global threshold T to the saliency of the points that consist S. Thresholding removes salient points with low saliency, that is,
$$S_T = \{(\boldsymbol{s}_i, \boldsymbol{v}_i, y_{D,i}) : y_{D,i} > T\}\,. \tag{6}$$
2. Select the point i in S_T with the highest saliency value and use it as a seed to initialize a salient region R_k. Add nearby points j to the region R_k as long as the intra-cluster variance does not exceed a threshold T_V. That is, as long as
$$\frac{1}{|R_k|}\sum_{j\in R_k} c_j^2 < T_V\,, \tag{7}$$
where R_k is the set of the points in the current region k and c_j is the Euclidean distance of the jth point from the seed point i.
3. If the overall saliency of the region R_k is lower than a saliency threshold T_S,
$$\sum_{j\in R_k} y_{D,j} \leq T_S\,, \tag{8}$$
discard the points in the region back to the initial set of points and continue from step 2 with the next highest salient point. Otherwise, calculate the Euclidean distance of the center of region R_k from the center of salient regions already defined, that is, from salient regions $R_{k'}, k' < k$.

4. If the distance is lower than the average scale of R_k, discard the points in R_k back to the initial set of points, and continue with the next highest salient point. Otherwise, accept R_k as a new cluster and store it as the mean scale and spatial location of the points in it.
5. Form a new set S_T consisting of the remaining salient points, increase the cluster index k and continue from step 2 with the next highest salient point.

By setting the threshold T_V in step 2, we define clusters that have local support and are well localized in space and time. In addition, we want to take the saliency of the points into consideration such that the overall saliency of the region is sufficient. We do this in step 3, by setting a saliency threshold, T_S. Finally, the purpose of step 4 is to accept clusters that are sufficiently distant from each other. To summarize, a new cluster is accepted only if it has sufficient local support, its overall saliency value is above the saliency threshold, and it is sufficiently distant in terms of Euclidean distance from already existing clusters.

3.3 Space-Time Warping

There is a large amount of variability between feature sets due to differences in the execution speed of the corresponding actions. Furthermore, we need to compensate for possible shifting of the representations forward or backward in time, caused by imprecise segmentation of the corresponding actions. To cope with both these issues, we propose a linear space-time warping technique with which we model variations in time using a time-scaling parameter a and a time-shifting parameter b. In addition, in order to achieve invariance against scaling, we introduce a scaling parameter c in the proposed warping technique. To accommodate this procedure, we propose the Chamfer distance as an appropriate distance measure, in order to cope with unequal number of features between different sets of salient points. More specifically, for two feature sets $F = \{(x_i, y_i, t_i), 1 \leq i \leq M\}$ and $F' = \{(x'_j, y'_j, t'_j), 1 \leq j \leq M'\}$ consisting of an M and M' number of features, respectively, the Chamfer distance of the set F from the set F' is defined as follows:

$$D(F, F') = \frac{1}{M} \sum_{i=1}^{M} \min_{j=1}^{M'} \sqrt{(x'_j - x_i)^2 + (y'_j - y_i)^2 + (t'_j - t_i)^2}\,. \tag{9}$$

From eq. 9 it is obvious that the selected distance measure is not symmetrical, as $D(F, F') \neq D(F', F)$. For recognition purposes, it is desirable to select a distance measure that is symmetrical. A measure that satisfies this requirement is the average of $D(F, F')$ and $D(F', F)$, that is,

$$D_c(F, F') = \frac{1}{2}(D(F, F') + D(F', F))\,. \tag{10}$$

Let us denote by $F_w = \{(cx_i, cy_i, at_i - b), 1 \leq i \leq M\}$ the feature set F with respect to feature set F'. Then, the distance between F' and F_w is given by eq. 9 as:

$$D(F_w, F') = \frac{1}{M} \sum_{i=1}^{M} \min_{j=1}^{M'} \sqrt{(x'_j - cx_i)^2 + (y'_j - cy_i)^2 + (t'_j - at_i + b)^2}\,. \tag{11}$$

Similarly, the feature set F' with respect to feature set F can be represented as $F'_w = \{(\frac{1}{c}x'_j, \frac{1}{c}y'_j, \frac{1}{a}t'_j + b), 1 \leq j \leq M'\}$ and their distance as:

$$D(F'_w, F) = \frac{1}{M'} \sum_{j=1}^{M'} \min_{i=1}^{M} \sqrt{(x_i - \frac{1}{c}x'_j)^2 + (y_i - \frac{1}{c}y'_j)^2 + (t_i - \frac{1}{a}t'_j - b)^2}\,. \tag{12}$$

The distance to be optimized follows from the substitution of eq. 11 and eq. 12 to eq. 10. We follow an iterative gradient descent approach for the adjustment of the a, b and c parameters. The update rules are given by:

$$a^{n+1} = a^n - \lambda_1 \frac{\partial D_c}{\partial a^n}, \tag{13}$$

$$b^{n+1} = b^n - \lambda_2 \frac{\partial D_c}{\partial b^n}, \tag{14}$$

$$c^{n+1} = c^n - \lambda_3 \frac{\partial D_c}{\partial c^n}\,, \tag{15}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the learning rates and n is the iteration index. The algorithm iteratively adjusts the values of a, b and c towards the minimization of the Chamfer distance between the two feature sets, given by eq. 10. The iterative procedure stops when the values of a, b and c do not change significantly or after a fixed number of iterations.

4 Tracking

4.1 Auxiliary Particle Filtering

Recently, particle filtering tracking schemes [10], [18], have been successfully used [58], [59], [19] in order to track the state of a temporal event given a set of noisy observations. Its ability to maintain simultaneously multiple solutions, called particles, makes it particularly attractive when the noise in the observations is not Gaussian and makes it robust to missing or inaccurate data.

The particle filtering tracking scheme described in this section is initialized at the spatiotemporal salient points that are detected using the procedure of section 3. Let c denote the template that contains the color information in a rectangular window centered at each detected salient point and α denote the unknown location of the facial feature at the current time instant. Furthermore, let us denote by $Y = \{y^1, \ldots, y^-, y\}$ the observations up to the current time instant. The main idea of the particle filtering is to maintain a particle based representation of the a posteriori probability $p(\alpha|Y)$ of the state α given all the observations Y up to the current time instance. The distribution $p(\alpha|Y)$ is represented by a set of pairs (s_k, π_k) such that if s_k is chosen with probability equal to π_k, then it is as if s_k was drawn from $p(\alpha|Y)$. Our knowledge about the a posteriori probability is updated in a recursive way. Suppose that we have a particle based representation of the density $p(\alpha^-|Y^-)$, that is we have a collection of K particles and their corresponding weights (i.e. (s_k^-, π_k^-)). Then, the Auxiliary Particle Filtering can be summarized as follows:

1. Propagate all particles s_k^- via the transition probability $p(\alpha|\alpha^-)$ in order to arrive at a collection of K particles μ_k.
2. Evaluate the likelihood associated with each particle μ_k, that is let $\lambda_k = p(y|\mu_k; c)$.
   For the definition of $p(y|\mu_k; c)$ we use, in this paper, the observation model described in [19].
3. Draw K particles s_k^- from the probability density that is represented by the collection $(s_k^-, \lambda_k \pi_k^-)$. In this way, the auxiliary particle filter favors particles with high λ_k, that is particles which, when propagated with the transition density, end up at areas with high likelihood.
4. Propagate each particle s_k^- with the transition probability $p(\alpha|\alpha^-)$ in order to arrive at a collection of K particles ${s_k}'$.
5. Assign a weight ${\pi_k}'$ to each particle as follows,

$$ {w_k}' = \frac{p(y|{s_k}'; c)}{\lambda_k}, \quad {\pi_k}' = \frac{{w_k}'}{\sum_j w_j} \tag{16} $$

   This results in a collection of K particles and their corresponding weights (i.e. $\{({s_k}', {\pi_k}')\}$) which is an approximation of the density $p(\alpha|Y)$.

4.2 Online Background Estimation

The particle filtering tracking scheme described in the previous section is initialized at the spatiotemporal salient points that are detected using the procedure described in section 3. As indicated from eq. 1, the input signal that is used is the convolution of the original image sequence with a Gaussian derivative filter along the temporal dimension. The result of this is that the detected salient points are localized on the edges of the moving objects existing in the scene, rather than on the objects themselves. This fact may deteriorate the output of the tracker used, since the patches of the sequence that are being tracked also include a considerable portion of the scene's background. For this reason, we implement the adaptive background estimation algorithm described in [54], in order to determine which pixels belong to the foreground and which ones to the background. According to this algorithm, the values of a particular pixel over time are considered as a temporal process. At each time t, what is known about a particular pixel (x_0, y_0) is its history:

$$ \{X_1, \ldots, X_t\} = \{I(x_0, y_0, i) : 1 \leq i \leq t\}, \tag{17} $$

where I is the image sequence. The recent history of each pixel is modeled by a mixture of K Gaussian distributions. The probability of observing the current pixel value is given by:

$$ P(X_t) = \sum_{i=1}^{K} w_{i,t} \cdot \eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \tag{18} $$

where K is the number of distributions, $w_{i,t}$ is an estimate of the weight of the i_{th} Gaussian in the mixture at time t, $\mu_{i,t}$ is the mean value of the i_{th} Gaussian

in the mixture at time t, $\Sigma_{i,t}$ is the covariance matrix of the i_{th} Gaussian in the mixture at time t, and η is a Gaussian probability density function. K was set to 3, and the covariance matrix Σ is assumed to be diagonal, meaning that the RGB values of the pixels are assumed to be uncorrelated.

The parameters of each Gaussian mixture were initially estimated using the Expectation-Maximization (EM) algorithm and by using a small portion of the available data (i.e. the first few frames of the image sequence). Subsequently at each new frame t we follow an update procedure similar to the one of [54]. Every new pixel value X_t is checked against the existing K distributions until a match is found. A match is defined if the current pixel is within 3 standard deviations of a distribution. In case a match is found the parameters of the Gaussians are updated. If none of the K distributions match the current pixel value, the least probable distribution is replaced with a distribution with the current value as its mean value, an initially high variance, and low prior weight.

At each iteration of the particle filtering tracking scheme of section 4.1, every new particle is evaluated based on an invariant colour distance between the initial template (centered at the initializing spatiotemporal salient point) and the block that corresponds to the particle that is being evaluated. In order to take the estimated background model into account, we add an additional cost in the evaluation process of each new particle. The additional cost for every pixel is equal to the probability that the pixel belongs to the current background model, that is,

$$C_{i,j,t} = \sum_{i=1}^{K} w_{i,j}\eta(X_{j,t}, \mu_{i,j,t}, \Sigma_{i,j,t}), \tag{19}$$

where K is the number of distributions, $w_{i,j,t}$ is an estimate of the weight of the i_{th} Gaussian in the mixture for the pixel j at time t, $\mu_{i,j,t}$ is the mean value

Fig. 2. Initial estimation of the background for an action where the subject is just raising its right hand

of the i_{th} Gaussian in the mixture for the pixel j at time t and $\Sigma_{i,j,t}$ is the covariance matrix of the i_{th} Gaussian in the mixture for pixel j at time t.

If a pixel in the block belongs to the background, then eq. 19 will assign a large cost to that pixel, since the resulting probability will be high. If most pixels in the block belong to the background, then the additional cost to that block will also be large and consequently, a smaller weight will be assigned to it by the particle filter. In this way, the tracking scheme favors blocks that contain larger number of foreground pixels and assigns larger weights to the corresponding particles.

In Fig. 2 the initial background model that was estimated for an action where the subject is raising its right hand is presented. As can be seen from the figure, parts of the body that do not present significant motion are also considered part of the background. On the other hand, fast moving parts (e.g. right hand) are considered to belong to the foreground and are not included in the estimation.

5 Recognition

5.1 Longest Common Subsequence (LCSS) Algorithm

Using the analysis of the previous sections, we represent a given image sequence by a set of short trajectories, where each trajectory is initialized at a point which is considered salient both in space and time. Formally, an image sequence is represented by a set of trajectories $\{A_i\}, i = 1 \dots K$, where K is the number of trajectories that consist the set. Each trajectory is defined as $A_i = ((t_{i,n}, x_{i,n}, y_{i,n}), \dots)$, $n = 1 \dots N$, where $t_{i,n}, x_{i,n}, y_{i,n}$ are spatiotemporal coordinates and N is the number of samples that consist A_i. Let us define another trajectory set $\{B_j\}$, $j = 1 \dots L$ representing a different image sequence. Similar to $\{A_i\}$, the trajectories in $\{B_j\}$ are defined as $B_j = ((t_{j,m}, x_{j,m}, y_{j,m}), \dots)$, $m = 1 \dots M$, where M is the number of individual trajectories that consist $\{B_j\}$. We use a variant of the LCSS algorithm presented at [55], [56] in order to compare the two sets. Before we proceed with the comparison, we align the two sets in space and time using the a, b and c parameters that were computed using the procedure of section 3.3. Let us define the function $Head(A_i) = ((t_{i,n}, x_{i,n}, y_{i,n})), n = 1 \dots N-1$, that is, the individual trajectory A_i reduced by one sample. Then, according to the LCSS algorithm, the distance between individual trajectories A_i and B_j is given by:

$$d_L(A_i, B_j) = \begin{cases} 0, \quad if\ A_i\ or\ B_j\ is\ empty \\ \\ d_e((t_{i,n}, x_{i,n}, y_{i,n}), (t_{j,m}, x_{j,m}, y_{j,m})) \\ +d_L(Head(A_i), Head(B_j)), \\ if\ \ |t_{i,n} - t_{j,m}| < \delta\ and\ \ |x_{i,n} - x_{j,m}| < \varepsilon \\ and\ \ |y_{i,n} - y_{j,m}| < \varepsilon \\ \\ \max(d_L(Head(A_i), B_j), d_L(A_i, Head(B_j))) + p, \\ otherwise \end{cases} , \tag{20}$$

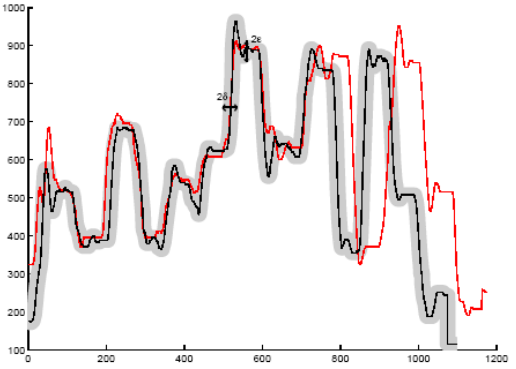


Fig. 3. The notion of the LCSS matching within a region of δ and ϵ of a trajectory

where d_e is the Euclidean distance, δ controls how far in time we can go in order to match a given point from one trajectory to a point in another trajectory, ϵ is the matching threshold and p is a penalty cost in case of mismatch. The notion of the LCSS distance of eq. 20 is depicted in Fig. 3.

Subsequently, the distance between sets $\{A_i\}$ and $\{B_j\}$is defined as follows:

$$D_L(\{A_i\},\{B_j\}) = \frac{1}{K}\sum_i \min_j d_L(A_i,B_j) + \frac{1}{L}\sum_j \min_i d_L(B_j,A_i), \tag{21}$$

that is, the average over the set of the minimum distances, as they have been defined in eq. 20, between the K trajectories of set $\{A_i\}$ and the L trajectories of set $\{B_j\}$.

5.2 Relevance Vector Machine Classifier

We propose a classification scheme based on Relevance Vector Machines [57] in order to classify given examples of human actions. A Relevance Vector Machine (RVM) is a probabilistic sparse kernel model identical in functional form to the Support Vector Machines (SVM). In their simplest form, Relevance Vector Machines attempt to find a hyperplane defined as a weighted combination of a few Relevance Vectors that separate samples of two different classes. In contrast to SVM, predictions in RVM are probabilistic. Given a dataset of N input-target pairs $\{(F_n, l_n), 1 \leq n \leq N\}$, an RVM learns functional mappings of the form:

$$y(F) = \sum_{n=1}^{N} w_n K(F, F_n) + w_0\,, \tag{22}$$

where $\{w_n\}$ are the model weights and $K(.,.)$ is a Kernel function. Gaussian or Radial Basis Functions have been extensively used as kernels in RVM. In our case, we use as a kernel a Gaussian Radial Basis Function defined by the distance measure of eq. 21. That is,

$$K(F,F_n) = e^{-\frac{D_L(F,F_n)^2}{2\eta}}\,, \tag{23}$$

where η is the Kernel width. RVM performs classification by predicting the posterior probability of class membership given the input F. The posterior is given by wrapping eq. 22 in a sigmoid function, that is:

$$p(l|F) = \frac{1}{1 + e^{-y(F)}} \tag{24}$$

In the two class problem, a sample F is classified to the class $l \in [0, 1]$, that maximizes the conditional probability $p(l|F)$. For L different classes, L different classifiers are trained and a given example F is classified to the class for which the conditional distribution $p_i(l|F)$,$1 \leq i \leq L$ is maximized, that is:

$$Class(F) = \arg\max_i (p_i(l|F)) \,. \tag{25}$$

6 Experimental Results

For the evaluation of the proposed method, we use aerobic exercises as a test domain. Our dataset consists of 12 different aerobic exercises, performed by amateurs, that have seen a video with an instructor performing the same set of exercises. Each exercise is performed twice by four different subjects, leading to a set of 96 corresponding feature sets.

In order to illustrate the ability of the proposed method to extract the kind of motion performed, we present in Fig. 4 the trajectories that were extracted

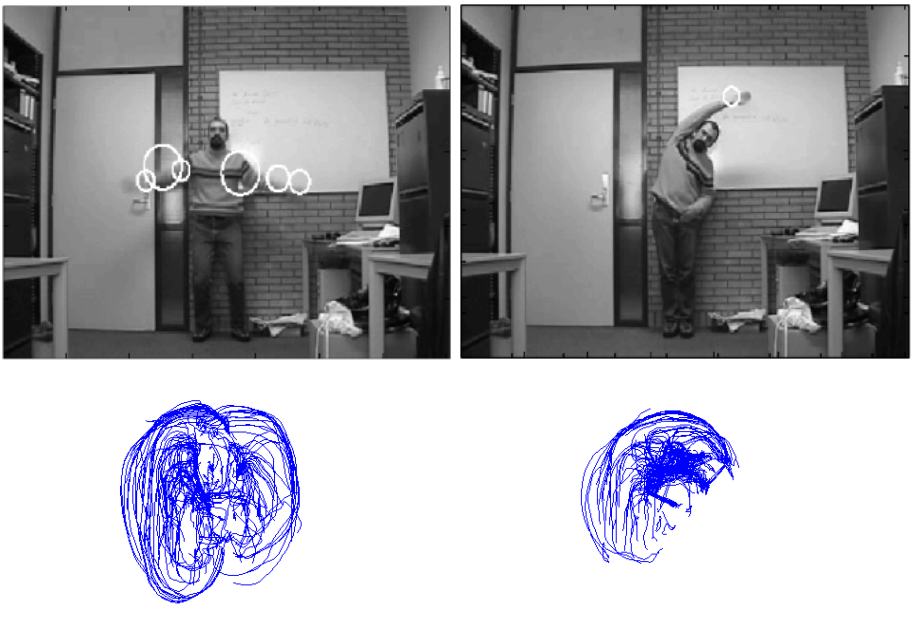


Fig. 4. Extracted trajectories for two different actions

Table 1. Recall and Precision rates for the kNN and RVM classifiers

Class Labels	**1**	**2**	**3**	**4**	**5**	**6**
RVM Recall	1	1	1	1	0.5	0.5
RVM Precision	1	1	1	1	0.44	0.4
Class Labels	**7**	**8**	**9**	**10**	**11**	**12**
RVM Recall	1	0.88	0.63	0.63	0.88	1
RVM Precision	1	1	0.63	0.83	0.88	1

Table 2. RVM Confusion Matrix

Class labels	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
1	**8**	0	0	0	0	0	0	0	0	0	0	0	8
2	0	**8**	0	0	0	0	0	0	0	0	0	0	8
3	0	0	**8**	0	0	0	0	0	0	0	0	0	8
4	0	0	0	**8**	0	0	0	0	0	0	0	0	8
5	0	0	0	0	**4**	5	0	0	0	0	0	0	9
6	0	0	0	0	4	**3**	0	0	3	0	0	0	10
7	0	0	0	0	0	0	**8**	0	0	0	0	0	8
8	0	0	0	0	0	0	0	**7**	0	0	0	0	7
9	0	0	0	0	0	0	0	1	**5**	2	0	0	8
10	0	0	0	0	0	0	0	0	0	**5**	1	0	6
11	0	0	0	0	0	0	0	0	0	1	**7**	0	8
12	0	0	0	0	0	0	0	0	0	0	0	**8**	8
Total	8	8	8	8	8	8	8	8	8	8	8	8	8

from two different actions along with a snapshot of the corresponding actions. The salient points that are visible in the upper part of the figure were used in order to extract some of the trajectories presented in the lower part of the same Figure. Furthermore, the extracted trajectory set seems to correctly capture the pattern of the motion performed. This can easily be observed from the arch-like trajectories of the lower part of the figure, which correspond to the motion of the subjects' hands.

In order to classify a test example using the Relevance Vector Machines, we constructed 12 different classifiers, one for each class, and we calculated for each test example F the conditional probability $p_i(l|F), 1 \leq i \leq 12$. Each example was assigned to the class for which the corresponding classifier provided the maximum conditional probability, as depicted in eq. 25. Note that for estimating each of the $p_i(l|F)$, an RVM is trained by leaving out the example F as well as all other instances of the same exercise that were performed by the subject from F. The corresponding recall and precision rates, calculated as an average of all test trials, are given in Table 1. The total recognition rate is equal to 80.61%, which is a relatively good performance, given the small number of examples with respect to the number of classes, and the fact that the subjects were not trained. In Table 2 the confusion matrix generated by the RVM classifier is also given.

The confusion matrix in Table 2 conceals the fact that for some of the misclassified examples the probability assigned by the RVM classifier to the correct matching move might be very close to the probability assigned to the move actually selected by the classifier. We used the average ranking percentile in order to extract this kind of information and to measure the overall matching quality of our proposed algorithm. Let us denote with r^{F_n} the position of the correct match for the test example $F_n, n = 1 \ldots N_2$, in the ordered list of N_1 match

values. Rank r^{F_n} ranges from $r = 1$ for a perfect match to $r = N_1$ for the worst possible match. Then, the average ranking percentile is calculated as follows:

$$\overline{r} = \left(\frac{1}{N_2} \sum_{n=1}^{N_2} \frac{N_1 - r^{F_n}}{N_1 - 1} \right) 100\%. \tag{26}$$

Since our dataset consists of 96 test image sequences divided in 12 separate classes, it follows that $N_1 = 12$ and $N_2 = 96$. Each of the 12 match values are provided for each example by the 12 trained RVM classifiers. The average ranking percentile for the RVM classifier is 94.5%. Its high value shows that for the majority of the missclassified examples, the correct matches are located in the first positions in the ordered list of match values.

7 Conclusions

In this work, previous work on spatiotemporal saliency was enhanced in order to extract a number of short trajectories from given image sequences. Each detected spatiotemporal point was used in order to initialize a tracker based on auxiliary particle filtering. A background estimation model was also implemented and incorporated into the particle evaluation process, in order to deal with inadequate localization of the initialization points and to improve, thus, the performance of the tracker. A variant of the LCSS algorithm was used in order to compare different sets of trajectories. The derived LCSS distance was used in order to define a kernel for the RVM classifier that was used for recognition. We have illustrated the efficiency of our representation in recognizing human actions using as a test domain aerobic exercises. Finally, we presented results on real image sequences that illustrate the consistency in the spatiotemporal localization and scale selection of the proposed method.

Acknowledgements

This work has been partially supported by the Dutch-French Van Gogh program (project VGP-62-604) and the work of A. Oikonomopoulos has been supported by the Greek State Scholarship Foundation (IKY). The data set was collected while I. Patras was with the ISIS group at the University of Amsterdam.

References

1. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. International Conference on Multimodal Interfaces (2006)
2. Wang, J.J., Singh, S.: Video analysis of human dynamics - A survey. Real Time Imaging **9** (2003) 321 – 346
3. Wang, L., Hu, W., Tan, T.: Recent Developments in Human Motion Analysis. Pattern Recognition **36** (2003) 585 – 601

4. Bar-Shalom, Y., Fortmann, T.: Tracking and Data Association. Academic Press (1988)
5. Julier, S., Uhlmann, J.: Unscented filtering and nonlinear estimation. Proceedings of the IEEE **92** (2004) 401–422
6. Wu, Y., Hu, D., Wu, M., Hu, X.: Unscented kalman filtering for additive noise case: Augmented versus nonaugmented. IEEE Signal Processing Letters **12** (2005) 357–360
7. LaViola, J.: A comparison of unscented and extended Kalman filtering for estimating quaternion motion. Proceedings of the American Control Conference **3** (2003) 2435 – 2440
8. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. Pattern Analysis and Machine Intelligence **27** (2005) 699 – 714
9. Gu, H., Ji, Q.: Information extraction from image sequences of real-world facial expressions. Machine Vision and Applications **16** (2005) 105 – 115
10. Isard, M., Blake, A.: Condensation conditional density propagation for visual tracking. International Journal of Computer Vision **29** (1998) 5 – 28
11. Isard, M., Blake, A.: Icondensation: Unifying low-level and high-level tracking in a stochastic framework. European Conference on Computer Vision **29** (1998) 893 – 908
12. Lichtenauer, J., Hendriks, M.R.E.: Influence of the observation likelihood function on particle filtering performance in tracking applications. Automatic Face and Gesture Recognition (2004) 767– 772
13. Chang, C., Ansari, R., Khokhar, A.: Multiple object tracking with kernel particle filter. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **1** (2005) 566– 573
14. Schmidt, J., Fritsch, J., Kwolek, B.: Kernel particle filter for real-time 3D body tracking in monocular color images. Automatic Face and Gesture Recognition (2006) 567– 572
15. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-Based Object Tracking. IEEE Trans. Pattern Analysis and Machine Intelligence **25** (2003) 564–577
16. Yang, C., Duraiswami, R., Davis, L.: Fast multiple object tracking via a hierarchical particle filter. Proc. IEEE Int. Conf. Computer Vision **1** (2005) 212– 219
17. Shan, C., Wei, Y., Tan, T., Ojardias, F.: Real time hand tracking by combining particle filtering and mean shift. Automatic Face and Gesture Recognition **1** (2004) 669– 674
18. Pitt, M., Shephard, N.: Filtering via simulation: auxiliary particle filtering. J. American Statistical Association **94** (1999) 590 –
19. Patras, I., Pantic, M.: Tracking deformable motion. IEEE International Conference on Systems, Man and Cybernetics (2005) 1066 – 1071
20. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. Automatic Face and Gesture Recognition (2004) 97 – 102
21. Matthews, I., Ishikawa, T., Baker, S.: The template update problem. In Proceedings of the British Machine Vision Conference. (2003)
22. Jepson, A., Fleet, D., El-Maraghi, T.: Robust Online Appearance Models for Visual Tracking. IEEE Trans. Pattern Analysis and Machine Intelligence **25** (2003) 1296–1311
23. Avidan, S.: Support Vector Tracking. IEEE Trans. Pattern Analysis and Machine Intelligence **26** (2004) 1064–1072

24. Gavrila, D., Davis, L.: 3-D Model-Based Tracking of Humans in Action: A Multi-view Approach. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (1996) 73–80
25. MacCormick, J., Isard, M.: Partitioned Sampling, Articulated Objects and Interface-Quality Hand Tracking. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (2000) 3–19
26. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Model-Based Hand Tracking Using a Hierarchical Bayesian Filter. IEEE Trans. Pattern Analysis and Machine Intelligence **28** (2006) 1372–1384
27. Chang, W., Chen, C., Hung, Y.: Appearance-guided particle filtering for articulated hand tracking. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **1** (2005) 235– 242
28. Sigal, L., Bhatia, S., Roth, S., Black, M., M. Isard, M.: Tracking loose-limbed people. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **1** (2004) 421–428
29. Wu, Y., Hua, G., Yu, T.: Tracking articulated body by dynamic Markov network. Proc. IEEE Int. Conf. Computer Vision **2** (2003) 1094– 1101
30. Han, T., Ning, H., Huang, T.: Efficient Nonparametric Belief Propagation with Application to Articulated Body Tracking. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **1** (2006) 214– 221
31. Fei, H., Reid, I.: Probabilistic Tracking and Recognition of Non-Rigid Hand Motion. Int. Workshop on Analysis and Modeling of Faces and Gestures (2003) 60–67
32. Elgammal, A., Shet, V., Yacoob, Y., Davis, L.: Exemplar-based tracking and recognition of arm gestures. Proc. Int. Symposium on Image and Signal Processing and Analysis **2** (2003) 656–661
33. Nickel, K., Seemann, E., Stiefelhagen, R.: 3D-Tracking of Head and Hand for Pointing Gesture Recognition in a Human-Robot Interaction Scenario. Automatic Face and Gesture Recognition (2004) 565–570
34. Deutscher, J., Blake, A., North, B., Bascle, B.: Tracking through Singularities and discontinuities by random sampling. Proc. IEEE Int. Conf. Computer Vision **2** (1999) 1144–1149
35. Black, M., Jepson, A.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. International Journal of Computer Vision **26** (1998) 63–84
36. Kato, M., Y.W.Chen, G.Xu: Articulated Hand Tracking by PCA-ICA Approach. Automatic Face and Gesture Recognition (2006) 329– 334
37. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **2** (2000) 126–133
38. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of non-rigid objects using mean-shift. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **2** (2000) 142–149
39. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions. International Journal of Computer Vision **50** (2002) 203–226
40. Rao, C., Gritai, A., Shah, M., Syeda-Mahmood, T.: View-invariant alignment and matching of video sequences. Proc. IEEE Int. Conf. Computer Vision **2** (2003) 939–945
41. Gavrila, D.: The Visual Analysis of Human Movement: A Review. Comp. Vision, and Image Understanding **73** (1999) 82–92
42. Aggarwal, J., Cai, Q.: Human Motion Analysis: A Review. Comp. Vision, and Image Understanding **73** (1999) 428–440

43. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Proc. IEEE Int. Conf. Computer Vision **2** (2005) 1395 – 1402
44. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **2** (2001) 123 – 130
45. Song, Y., Goncalves, L., Perona, P.: Unsupervised Learning of Human Motion. IEEE Trans. Pattern Analysis and Machine Intelligence **25** (2003) 814–827
46. Fanti, C., Zelnik-Manor, L., Perona, P.: Hybrid Models for Human Motion Recognition. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition **1** (2005) 1166–1173
47. Feng, X., Perona, P.: Human action recognition by sequence of movelet codewords. Proc. Int. Symposium on 3D Data Processing Visualization and Transmission (2002) 105–115
48. Figueroa, P., Leitey, N., Barros, R., Brenzikofer, R.: Tracking markers for human motion analysis. Proc. of IX European Signal Processing Conf., Rhodes, Greece (1998) 941 – 944
49. Moeslund, T., Nrgaard, L.: A brief overview of hand gestures used in wearable human computer interfaces. Technical Report CVMT 03-02, ISSN 1601-3646 (2003)
50. Haralick, R., Shapiro, L.: Computer and Robot Vision II. Addison-Wesley (1993) Reading, MA.
51. Gilles, S.: Robust Description and Matching of Images. PhD thesis, University of Oxford (1998)
52. Kadir, T., Brady, M.: Scale saliency: a novel approach to salient feature and scale selection. International Conference on Visual Information Engineering (2000) 25 – 28
53. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal Salient Points for Visual Recognition of Human Actions. IEEE Trans. Systems, Man and Cybernetics Part B **36** (2005) 710 – 719
54. Stauffer, C.: Adaptive background mixture models for real-time tracking. Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (1999) 246 – 252
55. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. Proc. International Conference on Data Engineering (2002) 673 – 684
56. Buzan, D., Sclaroff, S., Kollios, G.: Extraction and clustering of motion trajectories in video. Proceedings, International Conference on Pattern Recognition **2** (2004) 521 – 524
57. Tipping, M.: The Relevance Vector Machine. Advances in Neural Information Processing Systems (1999) 652 – 658
58. Su, C., Zhuang, Y., Huang, L., Wu, F.: A two-step approach to multiple facial feature tracking: Temporal particle filter and spatial belief propagation. Proc. IEEE Intl Conf. on Automatic Face and Gesture Recognition (2004) 433 – 438
59. Pantic, M., Patras, I.: Dynamics of Facial Expressions-Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. IEEE Trans. Systems, Man and Cybernetics Part B **36** (2006) 433 – 449

Modelling the Communication Atmosphere: A Human Centered Multimedia Approach to Evaluate Communicative Situations

Tomasz M. Rutkowski[1] and Danilo P. Mandic[2]

[1] Laboratory for Advanced Brain Signal Processing
Brain Science Institute RIKEN
Saitama, Japan
tomek@brain.riken.jp
http://www.bsp.brain.riken.jp/∼tomek/
[2] Department of Electrical and Electronic Engineering
Imperial College of Science, Technology and Medicine,
London, United Kingdom
d.mandic@imperial.ac.uk
http://www.commsp.ee.ic.ac.uk/∼mandic/

Abstract. This chapter addresses the problem of multimodal analysis of human face–to–face communication. This is imporant since in the near future, smart environments equipped with multiple sensory systems will be able to sense the presence of humans and assess recognize their behaviours, actions, and emotional states. The main goal of the presented study is to develop models of communicative/interactive events in multimedia (audio and video), suitable for the analysis and subsequent incorporation within virtual reality environments. Interactive, environmental, and emotional characteristics of the communicators are estimated in order to define the communication event as one entity. This is achieved by putting together results obtained in social sciences and multimedia signal processing under one umbrella – the communication atmosphere analysis. Experiments based on real life recordings support the approach.

1 Evaluation of Human Communication

The notion of *communication atmosphere* is often used to describe the synergy of cues coming from speech, facial expression and body language of humans participating in conversation. Although intuitively clear, there is no precise definition of this term, although it commonly refers to a "conversation episode". This comes as no surprise, since in the same vain as several other important and intuitively clear concepts related to human behavior, communication atmosphere is easily felt but hard to define, and even more difficult to evaluate. Yet, precise quantification of this notion is a prerequisite to intelligent information processing of human communication. Communication atmosphere as a perceptual measure is somewhat similar to the concept of *affordance* proposed in [1], since it evaluates communicative situations based on their intellegibility (potential to be understood).

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 155–169, 2007.

This Chapter addresses some of the open issues related to the modelling of communication atmosphere; this is achieved based on the three dimensions which span human communication process:- the environmental, interactive and emotional component. From an information theoretic perspective, the analysis of a synergy of these elements is essential in order to create and recreate the "climate" of a meeting (communication atmosphere). This can then be easily modelled and understood by an independent observer.

This concept is not new, indeed this has been subject of many psychological studies, especially when dealing with body language issues. In [2] attention level was defined as *"the degree of clarity of an experience ranging from unconsciousness (total lack of awareness) to focal attention (vivid awareness)"*; the nature of communicative events can then be evaluated based on attention/awerness levels. In [3] the impact of information amount that impinges on students' sensory registers during lectures was studied. The authors reported that students' involvement in the learning process was decreased when their attention level was not high. Attention, hence, plays an important role when it comes to the selection of sensory information for an efficient understanding of communication, since it is directly related to degree of involvement of the participants or spectators. In the proposed model of communication atmosphere, three elements (dimensions) of communicative situations are defined, based on interactive (social) features, emotional (mental) characteristics of the communicators, and environmental (physical) features of the place where the event occurs.

The existing studies on communicative activity monitoring, for instance the study in [4], focus on the problem of automatic evaluation of the activity in the audio and visual channel, with applications e.g. in distance learning. Despite the techical completeness, this and other standard approaches are limited only to the evaluation of communicative interaction:- they do not consider other aspects of the communication atmosphere, such as environmental and emotional ones.

The proposed approach can be analysed within the so called "W5+" framework (*who, where, what, when, why, how*), for a comprehensive survey see [5]. There, the authors discussed new challenging areas of research towards creation of human-centered interactive environments. Within this framework, our approach aims at modelling the behavioural dynamics and interactivity of human communication.

The applications of the proposed approach, such as those in multimedia, require tracking of the features (elements) of human communication, these can then be used for evaluation from a behavioral (interaction) point of view.

We propose a signal processing and data fusion based approach to the modelling of communication atmosphere. This is achieved based on a qualitative evaluation of visual and auditory attention levels of the participants in conversation. Section 2 introduces our three–dimensional approach to the analysis of communication atmosphere. Section 3 further elaborates on the proposal the proposed three-dimensional communication space. Examples and experimental results based on real world recordings, together with critical discussion conclude the Chapter.

2 Communication Atmosphere

Crucial issues which underpin the paradigm of monitoring and modelling the communication atmosphere are related to:- i) the selection of features which best describe a communication event; ii) subsequent identification of the relationships among those factors. Our underlying aim is therefore to model:-

- the extent to which the audiovisual behavior (interactions) influences the overall climate of a meeting;
- the extent to which the parameters of the environment where the communication takes place (room, virtual environment) affects the actual communication climate and the efficiency of communication;
- the extent to which emotional states of the participants in communication influence the overal climate of a meeting.

These factors are integral parts of any machine learning model of communication atmosphere (Gestalt); the notion of "communication interactivity" then relates to the states representing attention and involvement levels of the participants. We here propose a signal processing framework to address these issues. This is achieved based on a model in which human communication is perceived a dynamical system which lives in the three-dimensional space spanned by the elements of communication atmosphere. These three dimensions of the communication atmosphere are:-

Environmental dimension which comprises the ambient conditions in which communication takes place. This dimension naturally reflects the fact that communication episodes are not conducted in a synthetic and noise free information "vacuum", but instead in a real–world environment which can influence significantly our perception. This, in turn, changes the climate of a communication episode, as our attention levels enhance or deteriorate (acoustic and visual noise, traffic, other people);

Communicative dimension is introduced with the idea to characterize the behavioral aspects of a communication episode. This dimension reflects the ability of participants to interact and is is closely related to the notion of *communication efficiency*, a qualitative measure of the communication process which is directly related to the attention level and dynamics of partitipants' involvement [6,7];

Emotional dimension is related to emotional states of the participants. This dimension is introduced in order to account for subtleties of human behavior expressing the emotional states, these largely influence say the climate of a meeting. It is intuitively clear emotional dimensions of the participants in a communication episode are either totally synchronised, or that emotions expressed by the "sender" are to a certain extent reflected (after a delay) by the "receiver" [8,9]. These coupled emotional states of participants help create the overall climate of a communication event.

It is important to realise that human communication comprises both verbal and non–verbal elements, and that although both contribute to our understanding

of a communication event (Gestalt), their respective contribution to our assessment of the climate of human communication is dramatically different. This also relates to the information theoretic aspects of the audio and visual sensors in humans. The capacity of nonverbal communication was studied in [10], this was based on kinetics related features. According to this study, humans can differentiate between about 250,000 different facial expressions. This truly fascinating ability of humans poses a problem when it comes to automated classification of human expressions, since intelligent classifiers can only handle a much smaller number different patterns [11,12]. In [13], it was found that the verbal component of face–to–face communication events makes as little as 35% of the overal communication contents, whereas up to 65% of the actual communication can be conducted in a non-verbal manner. Experiments reported in [9] demonstrate that in business related communication episodes, body language accounts for between 60% and 80% of the overall impact.

Based on these observation, we shall focus our attention on a dynamical analysis of nonverbal components in human communication. Our working hypothesis is that the understanding of the dynamics of nonverbal communication would enable us to estimate the climate of the multimodal human communication. This will also provide a basis upon which other modalities of human communication can be used in order to provide a "greater picture".

2.1 Environmental Dimension

Notice that environment conditions not only influence our perception about objects and situations, but also we can change e.g. the level of external audio or video noise only to a limited extent [14]. We therefore propose to incorporate these environmental characteristics within our model. Physical features of the environment can be extracted after separation of the recorded information streams into two categories: i) items related to the communication process and ii) unrelated items (unsignificant) [7]. The general idea here is to separate the information from audio and video streams into background noise and useful signals [15,16].

To achieve this we first detect the presence of auditory and visual events that occur in the surrounding space but are not related to the communicators' actions (i.e. background audio and video). In our approach, the analysis of the environmental dimension is performed in two stages:-

- Estimation of power levels of noise and audio events not related to speech;
- Estimation of visual activity not related to participants in communication (background video).

The amount of the background audio energy (treated as noise) is estimated as a segmental signal-to-interference-ratio (SIR) since calculation of an integral signal-to-noise-ratio would not reflect temporal fluctuations of the situation dynamics (e.g. when communicators speak louder). The segmental auditory SIR, A_{SIR}, is calculated as:

$$A_{SIR}(m) = 10 \log_{10} \frac{\sum_{n=Nm}^{Nm+N-1} s_e^2(n)}{\sum_{n=nM}^{Nm+N-1} s_n^2(n)}, \quad (1)$$

where $s_n(n) = s_o(n) - s_e(n)$ is the noise estimate; $s_o(n)$ is the recorded audio signal (with noise); N stands for the number of audio samples; M is the number of speech segments.

Within the estimation of visual background "noise", we compare the activity in the communicators' areas with thoat in the background. The amount of visual flow is calculated as interference–like coefficient, V_{SIR}, given by:

$$V_{SIR}(m) = 10 \log_{10} \frac{\sum_{n=Nm}^{Nm+N-1} v_h(n)^2}{\sum_{n=nM}^{Nm+N-1} v_b(n)^2}, \quad (2)$$

where $v_b(n)$ represents visual flow features of the background and $v_h(n)$ are related to the extracted motion features of active communicators. Both A_{SIR}

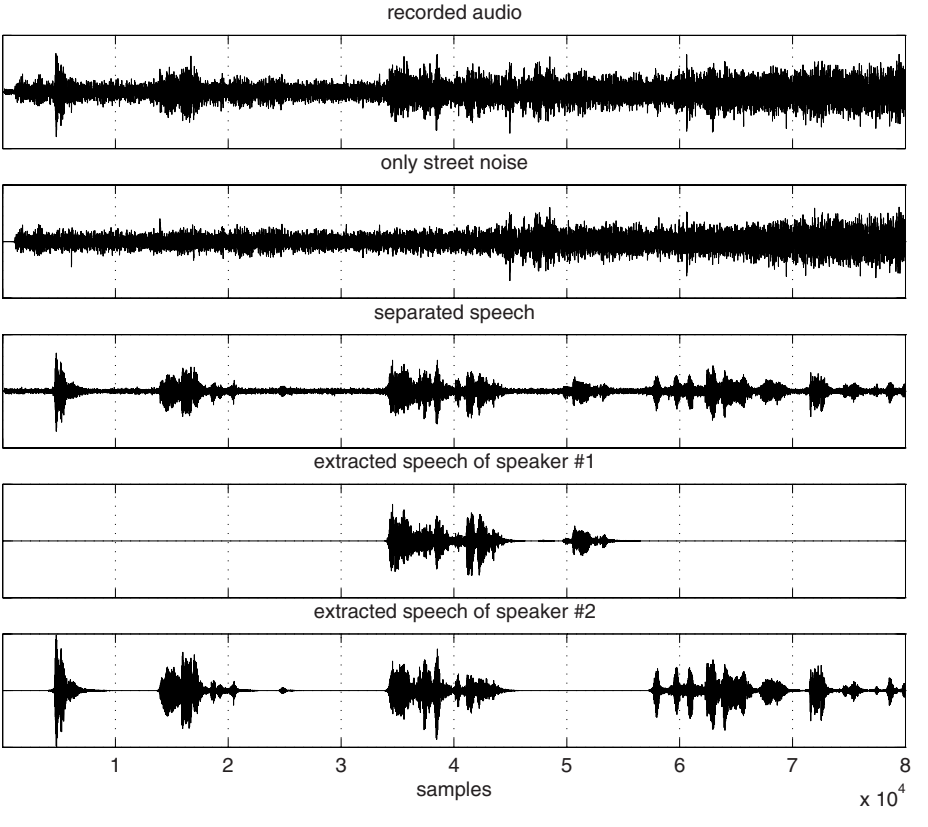


Fig. 1. Wide-band noise masks completely the mixture speech coming from two speakers (compare recorded signal in top box, only interference in second box from the top, enhanced speech in middle box and finally separated signals from two speakers - there is clear need for blind source separation (BSS))

and V_{SIR} are then combined to form a joint audiovisual SIR measure which characterises environmental conditions.

Figure 1 illustrates the procedure to evaluate the environmental dimension; the audio and video component are filtered from the total information related to the communication [15,16,17].

2.2 Emotional Dimension

Present approaches to emotional state estimation usually perform a static analysis of facial expressions and seldom consider a dynamic or multimodal analysis. Experiments reported in [8] and [9] showed that unconscious mind exerts direct control of facial muscles. It was demonstrated that facial emotions presented by a sender (e.g. a smile) were reciprocated by returning a smile by the receiver. The experiments reported in [8] were conducted using electromyography (EMG), so the the electrical activity of facial muscles could be captured. Results of those experiments revealed that the communicators often had no total control over their facial emotions. These findings suggest that important aspects of "emotional face–to–face" communication could occur at the unconscious level.

In our approach we estimate emotional states from available nonverbal features only, these actually reflect (partially) the psychological state of the communicator. Alternatively, emotional states of the communicators can be estimated from speech only [18], since the video features are harder to process and understand. At present emotions from speech are estimated based on three features:- the voice fundamental frequency with duration aspects related to stable preiods; the speech harmonic content; and the speech signal energy expressed as an averaged root mean square (RMS) value in voiced frames [7]. Primary human emotions, such as neutral, sad, angry, happy and joyful, which the communicator might experience during the communication event, can then be determined using machine learning algorithms [19].

2.3 Communicative Dimension

The third and, probably, most important component of the communication atmosphere model is the communicative dimension. It refers to the communicators' audiovisual behavior in terms of their ability to interact in a meaningful way during the conversation. The synchronization and interaction measures (efficiency-like) have been developed by the authors in their previous research [6,7]. The communication model used (hybrid linear/transactional) is linear in short time windows [20]. The active (in short time windows) communicator - the sender - is expected to generate more emphasised audiovisual flow, with breaks in their activity when the receiver responds. On the other hand, the passive (in short time windows) communicator – the receiver – is expected to react accordingly, by not disturbing (overlapping with) the sender's communication activity. Turn-taking (role changing) between the senders and receivers is a critical assumption in the hybrid communication model. Only the case of intentional communication, which occurs when all the communicators are willing to interact, is considered

here. All the situations when so-called metacommunication [20] occurs are out of the scope of this chapter.

In [6] the communication efficiency is defined as a measure that characterises behavioral coordination of the paraticipans in communication. This measure describes the communication process from the point of view of the interactivity between sending and receiving of information, as observed through the audiovisual channel. There is no means, however, to evaluate the understanding of the messages by the communicators, but it is assumed that some sort of feedback or the receiver's reaction should be presented. This study proposes a measure of the communication efficiency which is as a combination of four mutual information estimates between two visual (V_i), two audio (A_i), and two pairs of audiovisual features ($A_i; V_i$). First, the two mutual information estimates are evaluated for selected regions of interest (ROI), where the communicators may be present and their speech signal is active, as follows:

$$I_{A_i,V_i} = \frac{1}{2} \log \frac{|R_{A_i}|\,|R_{V_i}|}{|R_{A_i,V_i}|}, \tag{3}$$

where $i = 1, 2$, and R_{A_i}, R_{V_i}, R_{A_i,V_i} stand for empirical estimates of the respective covariance matrices of the feature vectors [6]. Next, the two mutual information estimates indicating simultaneous activity in the same modes (audio and video respectively) are calculated for video streams as:

$$I_{V_1,V_2} = \frac{1}{2} \log \frac{|R_{V_1}|\,|R_{V_2}|}{|R_{V_1,V_2}|}, \tag{4}$$

and, analogously, for audio streams as:

$$I_{A_1,A_2} = \frac{1}{2} \log \frac{|R_{A_1}|\,|R_{A_2}|}{|R_{A_1,A_2}|}. \tag{5}$$

where R_{A_1,A_2} and R_{V_1,V_2} are the empirical estimates of the respective covariance matrices for unimodal feature sets corresponding to different communicator activities. A conceptual sketch of this idea is shown in Figure 2. Quantities I_{A_1,V_1} and I_{A_2,V_2} evaluate the local synchronism between the audio (speech) and visual flows of the observed communicators (e.g. facial expressions). It is expected that the sender should have a higher level of synchronism, thus reflecting the higher activity. Quantities I_{V_1,V_2} and I_{A_1,A_2} are introduced to detect possible crosstalks in same modalities (audio-audio, video-video). The latter pair is also useful to detect possible overlappings in activities, these have a negative impact on the intelligibility of communication.

The communicator role (i.e. sender or receiver) can be estimated from the audiovisual mutual information features, which are extracted and monitored over time. It is assumed that a higher synchronisation across the audio and video features characterises the active member - the sender, while the lower synchronisation characterises the receiver. This implies that in efficient communication, we should have synchronised audiovisual behavior of the sender and unsynchronised behavior of the receiver. From the interacttions in audiovisual streams,

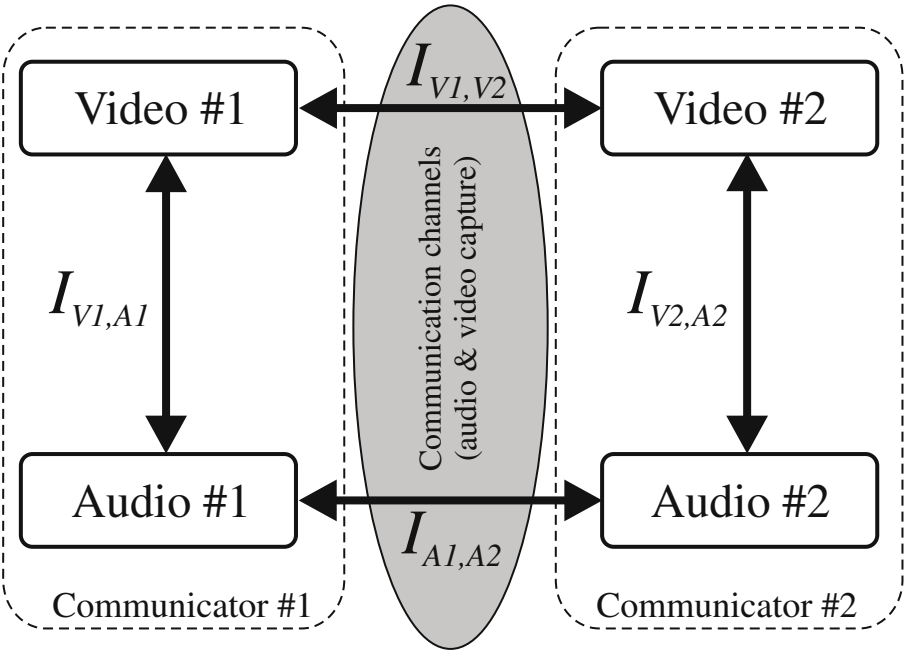


Fig. 2. Scheme for the communication synchronicity evaluation. Mutual information estimates I_{A_1,V_1} and I_{A_2,V_2} between audio and visual features streams of localized communicators account for the local synchronization. The estimates I_{A_1,A_2} and I_{V_1,V_2} detect crosstalks within the same modality.

the communicators' roles are classified as short-time senders and receivers [6,20]. As stated above, the efficient sender-receiver interaction during communication should involve both action and feedback. The pair of the mutual information estimates for the local synchronisation of the senders and the receivers in equation (3) is used to give clues about concurrent individual activities during the communication event, while the unimodal cross-activities estimates from equations (4) and (5) are used to evaluate the interlaced activities for a further classification.

There are many state-of-the-art techniques that attempt to solve the problem of recognition of communicative activities in particular modalities (e.g. audio only or video only), yet human communication involves interlaced verbal and nonverbal clues that constitute efficient or inefficient communication situations. For every communication situation the members are classified as a "sender", "receiver", or "in transition". In the presented approach, the interactions between individual participants is modelled from a stream of sequences of measurements, these are classified into streams of recognized phases using a machine learning algorithm [21]. A multistage and multisensory classification engine based on the linear support vector machines (SVM) approach in one-versus-rest-fashion is used to identify the phases during ongoing communication, based on the mutual information estimates from equations (3), (4), and (5). Results of such

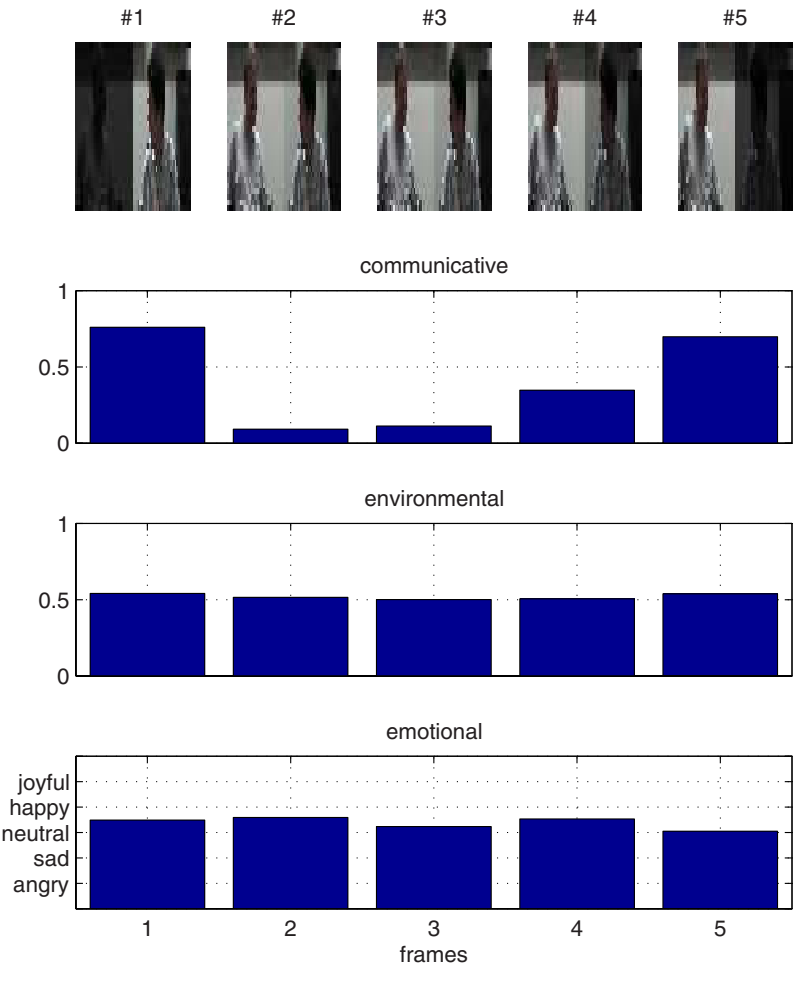


Fig. 3. Communication track along the three communicative dimensions. This short sequence started when right hand side communicator was a sender, then during three frames both were active, and finally left hand side communicator became the sender.

classification during an ongoing conversation are depicted in the form of shaded windows in Figure 3. The dark shades identify the receiver, while the light shades represent the sender. The transient phases are represented with the same shading for both communicators.

The communication efficiency, as proposed in [6], is a normalized value calculated from the integration over time of the evaluations of communicator's role. It reflects the analysis of the phases in which the communication is present (recognized the sender and receiver, or transient phases), together with the presence of possible crosstalks in audio and video channels (enhanced or unsynchronised receiver activities). This normalized measure is set to unity for smooth and

interlaced events, and its value is zero for completely overlapping communication activities.

The communicative efficiency is therefore estimated as follows:

$$C(t) = \left(1 - \frac{I_{V_1V_2}(t) + I_{A_1A_2}(t)}{2}\right) \cdot \cdot \left|I_{A_1V_1}(t) - I_{A_2V_2}(t)\right|, \tag{6}$$

and it takes values in the range $0 \leq C(t) \leq 1$, since all the mutual information estimate values are limited to $\langle 0, 1 \rangle$.

3 Tracking the Communication Atmosphere

Communication atmosphere, as defined in this study, is a region in the three-dimensional space (see Figure 4), obtained by independently estimating the environmental impact, the communication efficiency and the communicators' emotions in the ongoing communication process. This measure allows for the communication process evaluation and adjustment.

The communication atmosphere definition can be formalized as:

$$A(t) = A\left(E(t), C(t), M(t)\right), \tag{7}$$

where $A(t)$ represents the communication atmosphere evaluation at a time t; it is a function of the environmental estimate E, the communication level estimate C (also the communication efficiency measure), and the communicators' emotion estimate M.

The estimate $A(t)$ characterises communication atmosphere at a given time. One less sensitive measure would be the average for a given time window:

$$A_{avg(t_a,t_b)} = \frac{1}{|t_b - t_a|} \sum_{t=t_a}^{t_b} A(t), \tag{8}$$

where $t_a > t_b$, and short time functions:

$$A_{t_a,t_b} = \{A(t_a), \ldots, A(t_b)\}, \tag{9}$$

The short time trajectories A_{t_a,t_b} in the three-dimensional space can later be used as inputs for situation classifiers or three-dimensional communication climate models. This problem belongs to the area of communicative semantics, where multiple sequentially structured simultaneous communication processes are in a dialogical relationship. Within such models, the particular focus should be on beats and deixis [2], as lower-level action structure the foregrounding and backgrounding of the higher-level actions that participants are simultaneously engaged in. The problem of classification of communicative semantics–related stages is a subject for our future research, this will explore the ultimate relation between the communication atmosphere and communicative situation affordances or communicative situation norms as proposed by [22] and modeled

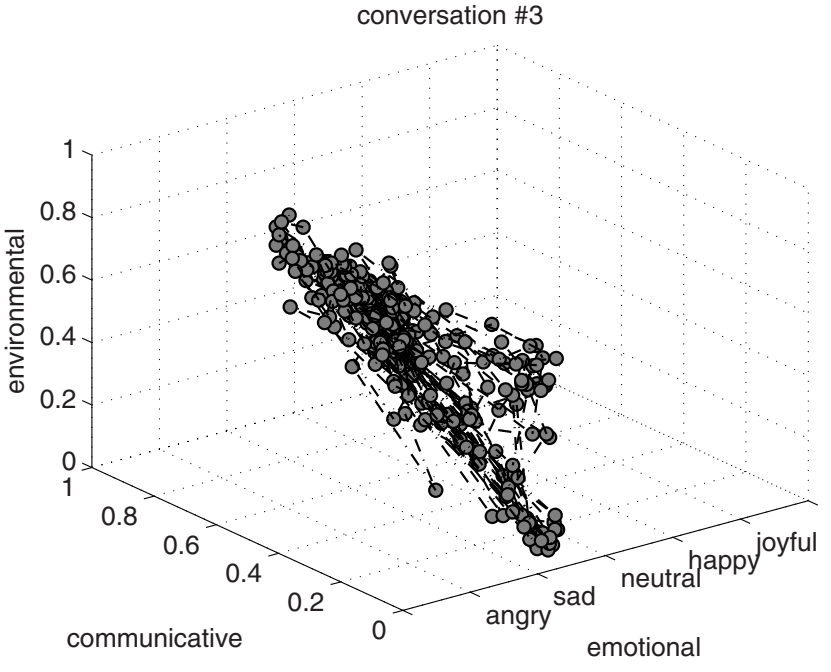


Fig. 4. The communication atmosphere space spanned among three dimensions. The trajectory represents a face–to–face communication situation in a relatively quiet environment.

by [23]. The plot of short time functions A_{t_a,t_b} is presented in Figure 4 as a three-dimensional trajectory, and in Figure 3 – as the three separated bar plots for every dimension showing a vivid independence in time among the chosen dimensions.

3.1 The Atmosphere Adjustment

Once the communication atmosphere features are estimated, it is possible to manipulate their values and appropriately postprocess the recorded multimedia content to obtain the desired climate. It is possible to independently manipulate characteristics of the environmental dimension by increasing or decreasing the auditory or visual presence of the environment. The information separation discussed in previous sections can be used in the opposite way to add or remove the environmental components. In a similar way, it is possible to edit emotional features of the communicators' auditory activities in order to change the emotional component of the overall climate of recorded communication. The communicative dimension can also be modified by adjusting the occurrences of communicators' interactions (e.g. by adding or removing silent breaks) in time. An example of the original and adjusted communication atmosphere tracks (only for environmental and emotional dimensions in this case) is presented in Figure 5.

4 Experimental Results

The approach presented in this study was tested in two experiments. In the first experiment, two sets of cameras and microphones where used to capture ongoing communication events. Two pairs of synchronized video and stereophonic audio streams were recorded. In the second experiment, we utilized a single high definition digital video camera (HDV) with a stereo microphone. This setup is similar to usual video recordings broadcasted in television channels. Both setups allowed capture of facial areas with higher resolution, which are highly synchronized with speech waveforms [6,7]. In both experiments conducted in laboratory controlled environments, the subjects (communicators) were asked to talk freely in face–to–face situations. We focused on the interlaced communication analysis, so that the subjects were asked to make a conversation with frequent turn taking (the discussion style). Such instruction given to subjects had a side effect of increased attention, which had positive impact on our assumption of intentional communication analysis. The experiments were conducted to validate the thesis, that the separate analysis of the three dimensions related to communication can be performed and allows for comprehensively describing the process as a whole (feature independence). For the communicative dimension,

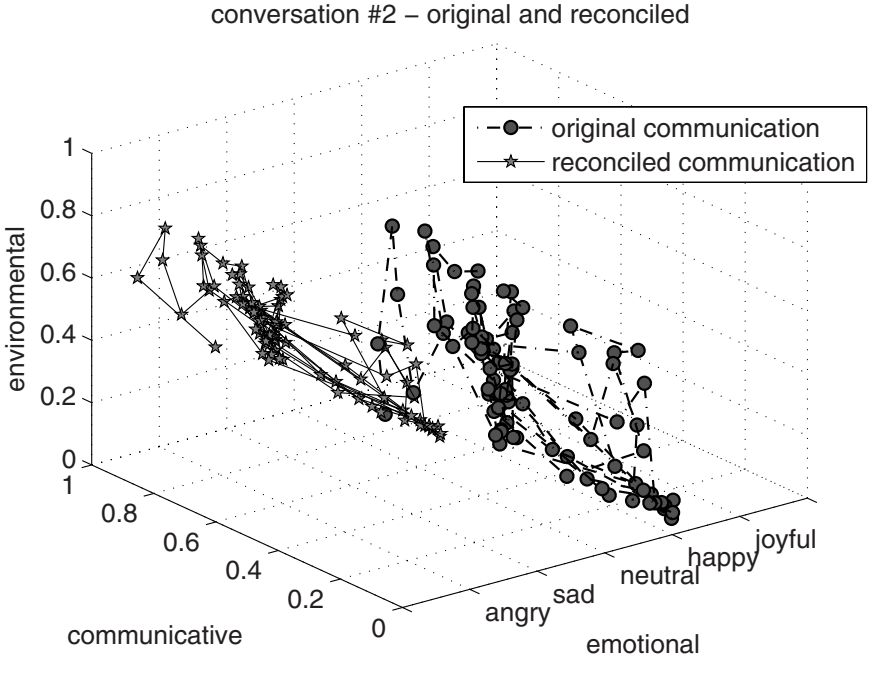


Fig. 5. Communication atmosphere adjustment: The presented trajectory of a face–to–face communication was modified in the environmental (subtraction of environmental noise) and emotional (shifting of the emotional features of the communicators' voices) dimensions

estimation of the communication efficiency is based on mutual information in multimedia streams. The track of the integrated communication efficiency value over an ongoing person-to-person communication event is shown in Figure 3. The normalized values close to one indicate the moments when the interaction was "proper", crosstalks in audio and visual channels (the local, negative synchronization) did not occur, and there was only single active person at each time (the lighter area around the person in the top box). Low values close to zero correspond to the transitional situations, or when both parties are active at the same time. The communicative situation analysis in this three-dimensional space allows for tracking of communication events and later classifying from obtained trajectories (compere the shapes of the trajectories of two independent events with different communicators but with similar discussion topics shown in Figures 4 and 5). The adjustment procedure was performed over the the recorded communication after the analysis. In the current approach any or all the communication atmosphere dimensions can be considered. An example of a manipulated atmosphere trajectories before and after adjustment is presented in Figure 5, where the environmental and mental dimensions were modified.

5 Conclusions

This study has proposed a novel framework suitable for the analysis of human communication in the context of "W5+", defined in [5]. The proposed three dimensional analysis environment helps bridge the gap between psychological studies on communication atmosphere and machine learning applications of this problem. The proposed approach has been shown to be flexible enough to allow for both the estimation and an arbitrary modification of the the climate of human communication. Based on mutual information within multimodal data streams we have proposed a framework to identify and to classify participants in human communication, according their role. In the current study, the three dimensions of the communication atmosphere model were considered independent. The dependencies between these are subject of our future work.

Acknowledgements

The authors would like to thank Prof. Toyoaki Nishida, Prof. Michihiko Minoh, and Prof. Koh Kakusho of Kyoto University for their support and fruitful discussions. The work of Dr Rutkowski was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Creative Scientific Research, 13GS0003. Dr Mandic was partially supported by the Royal Society grant RSRG 24543.

We are grateful to Prof. Victor V. Kryssanov from Ritsumeikan University in Kyoto for his input during the setup stage of this research. We have also benefited from our discussions with Dr. Anil Bharath from Imperial College London.

References

1. Gibson, J.J.: The theory of affordances. In Shaw, R., Bransford, J., eds.: Perceiving, Acting and Knowing. Erlbaum, Hillsdale, NJ (1977)
2. Norris, S.: Analyzing Multimodal Interaction - A Methodological Framework. Routledge (2004)
3. Moore, D.M., Burton, J.K., Myers, R.J.: Multiple-channel communication: The theoretical and research foundations of multimedia. In Jonassen, D., ed.: Handbook of Research for Educational Communications and Technology. Prentice Hall International (1996) 851–875
4. Chen, M.: Visualizing the pulse of a classroom. In: Proceedings of the Eleventh ACM International Conference on Multimedia, ACM Press (2003) 555–561
5. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. In: Proceedings of The ACM International Conference on Multimodal Interfaces. (2006) 239–248
6. Rutkowski, T.M., Seki, S., Yamakata, Y., Kakusho, K., Minoh, M.: Toward the human communication efficiency monitoring from captured audio and video media in real environments. In: Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2003, Part II. Volume 2774 of Lecture Notes in Computer Science., Oxford, UK, Springer-Verlag Heidelberg (September 3–5 2003) 1093–1100
7. Rutkowski, T.M., Kakusho, K., Kryssanov, V.V., Minoh, M.: Evaluation of the communication atmosphere. In: Proceedings of 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2004, Part I. Volume 3215 of Lecture Notes in Computer Science., Wellington, New Zealand, Springer-Verlag Heidelberg (September 20–25 2004) 364–370
8. Dimberg, U., Thunberg, E., Elmhed, K.: Unconscious facial reactions to emotional expressions. Psychological Science **11**(1) (2000) 149–182
9. Pease, A., Pease, B.: The definitive book of body language - How to read others' thoughts by their gestures. Pease International (2004)
10. Birdwhistell, R.: The language of the body: the natural environment of words. In: Human Communication: Theoretical Explorations. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, USA (1974) 203–220
11. Ralescu, A., Hartani, R.: Fuzzy modeling based approach to facial expressions understanding. Journal of Advanced Computational Intelligence **1**(1) (October 1997) 45–61
12. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (December 2000) 1424–1445
13. Mehrabian, M.: Silent Messages. Wadsworth, Belmont, California (1971)
14. Rutkowski, T.M., Cichocki, A., Barros, A.K.: Speech enhancement using adaptive filters and independent component analysis approach. In: Proceedings of International Conference on Artificial Intelligence in Science and Technology, AISAT2000, Hobart, Tasmania (December 17–20 2000) 191–196
15. Rutkowski, T.M., Yokoo, M., Mandic, D., Yagi, K., Kameda, Y., Kakusho, K., Minoh, M.: Identification and tracking of active speaker's position in noisy environments. In: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC2003), Kyoto, Japan (September 2003) 283–286
16. Liu, W., Mandic, D.P., Cichocki, A.: Blind second-order source extraction of instantaneous noisy mixtures. IEEE Transactions on Circuits and Systems II **53** (2006) 931–935

17. Liu, W., Mandic, D.P.: A normalised kurtosis based algorithm for blind source extraction from noisy measurements. Signal Processing **86** (2006) 1580–1585
18. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: Proceedings of The Fourth International Conference on Spoken Language Processing, ICSLP'96. Volume 3., Philadelphia, PA (1996) 1970–1973
19. Vapnik, V.: The Nature of Statistical Learnig Theory. Springer Verlag (1995)
20. Adler, R.B., Rodman, G.: Undestanding Human Communication. Oxford University Press (2003)
21. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks **13**(2) (2002)
22. Stamper, R.: Signs, information, norms and systems. In Holmqvist, B., Andersen, P., Klein, H., Posner, R., eds.: In Signs of Work. De Gruyter, Berlin (1996) 349–399
23. Kryssanov, V., Kakusho, K.: From semiotics of hypermedia to physics of semiosis: A view from system theory. Semiotica (2005) (in press).

Modeling Influence Between Experts

Wen Dong and Alex Pentland

E15-383, 20 Ames Street
The MIT Media Laboratory
Cambridge, MA, 02139-4307
{wdong, sandy}@media.mit.edu

Abstract. A common problem of ubiquitous sensor-network computing is combining evidence between multiple agents or experts. We demonstrate that the *latent structure influence model*, our novel formulation for combining evidence from multiple dynamic classification processes ("experts"), can achieve greater accuracy, efficiency, and robustness to data corruption than standard methods such as HMMs. It accomplishes this by simultaneously modeling the structure of interaction and the latent states.

1 Introduction

Human computing is the next generation human-computer interaction scheme [8]. In this scheme, a multitude of unobtrusive and ubiquitous sensors work with each other, intelligently sense human behavior and interaction, and provide assistance accordingly. The *human computing* paradigm is different from the traditional keyboard/mouse multimodal scheme: The traditional scheme is computer-centered, and requires considerable efforts from a human user in order to make a computer understand. In comparison, the human computing scheme is human based. It requires the intelligent sensors to predict human behavior and interaction precisely without even being noticeable by a human user, in order to provide the best assistance possible.

Combining evidence from different sensors, classifiers, agents, or experts is a major algorithmic challenge in human computing. We propose the influence model as a new, efficient, and robust method for combining evidence from different dynamic processes. The influence model is analogous to the work of a team of experts. In this team, different experts cope with different types of data and use different statistical models. Each expert consults with the other experts about their results, instead of their raw data, to form a better understanding of the situation. The experts find others whose results have information about their particular problem, and form networks that can be used to pool data, identify outliers, etc.

The influence model has proven to be an efficient, robust method for analyzing multi-expert dynamics problems. It is in the tradition of N-heads dynamic programming on coupled hidden Markov models [7], the observable structure influence model [1], and the partially observable influence model [2], but extends these previous models by providing greater generality, accuracy, and efficiency.

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 170–189, 2007.

In this introductory section, we will first outline two types of several experimental platforms where we have used the influence modeling methodology. This will allow us to sketch some of the opportunities and challenges for our method of combining experts. We believe the same method can be applied to other human computing scenarios, and can be a good candidate for combining evidence of human physiological signals as well as social signals.

Example Applications

Our first example illustrating the general problem of combining expert classifiers is the *inSense* system [3] which combined several wearable sensor systems, each of which attempted to classify the state of the wearer. The ultimate goal of this system was to combine these local experts into a global estimate of wearer's state, and use this to control data collection by camera and microphone. The system was developed as part of the DARPA *Advanced Soldier Sensor Information System and Technology* (*ASSIST*) program [4].

In the *inSense* system (see Figure 1), data are collected by two accelerometers (worn on left hip and right wrist), an ambient audio recorder, and a camera (the latter two worn on the chest). From the data, four within-category "experts" are able to recognize various types of wearer context. These context categories include (1) eight types of locations (office, home, outdoors, indoors, restaurant, car, street, and shop), (2) six ambient audio configurations (no speech, user speaking, other speaking, distant voice, loud crowd, and laughter), (3) seven postures (unknown, lying, sitting, standing, walking, running, and biking), and (4) eight activities (no activity, eating, typing, shaking hands, clapping hands, driving, brushing teeth, and washing the dishes).

The within-category contexts recognized by the four category experts are then combined to determining moments of interest in the wearer's everyday life, which are then recorded as a sort of personal diary. The context estimates can also be used to assist a wearer in real time. Several things are worthy of mentioning in the *inSense* system. First, the contexts in the recent past provide information for recognizing the current contexts. This type of information can be utilized by, for instance, hidden Markov models.

Second, the four categories of contexts (postures, ambient audio configurations, postures, and activities) are related. For example, knowing that a user is typing would strongly bias the system to believe that he is both sitting and in his office / home, while knowing that a user is in his office / home only weakly hints that he is typing. Since the *inSense* system does not have a GPS, most of the user's location contexts are inferred from the other three categories of contexts.

Third, the relation between these four categories is too complex to be specified manually. The contexts from one category can be combined freely with those from another category, and there are as many as $8 \times 6 \times 7 \times 8 = 2688$ number of combined states. As a result, a good algorithm for this system must be able to explore and exploit the relations among different categories of contexts automatically while avoiding consideration of the exponential number of combined states.

Fig. 1. Left: the *inSense* system; Right: the *ASSIST* system

A very similar approach was taken in developing a soldier state recognition algorithm for the *ASSIST* system (see Figure 1), in collaboration with IBM and Georgia Tech teams. In this system, data was collected in real-time from several accelerometers, microphones, cameras, and a GPS/altimeter, all attached to different parts of the soldiers' clothing. Inference of soldier state was made in real-time, and data automatically shared among different soldiers wearing the *ASSIST* systems based on the pattern of activity shown among the group of soldiers.

Both systems used the same team-of-experts approach. In the *inSense* system, the raw data are computed upon in fundamentally different ways to get the four categories of contexts. The understandings in different categories of contexts can serve to enhance each other. In addition, there are so many combinations of contexts that the curse of dimensionality should be carefully avoided. In comparison, the sample sequences from different sensors in the *ASSIST* system observe very different probability laws and can hardly be jointly Gaussian, thus it was beneficial to apply different models to different sensors and then combine the results.

Since sensor failures were unavoidable due to insufficient power supply, sensor faults, connection errors, or other unpredictable causes, the team-of-experts approach allowed us to offset deficiencies in sensor data by providing estimates of the missing data that were constructed from the other experts' results.

A second example illustrates our approach from the problem of combining several smart sensors on a single individual to the problem of modeling a group of persons over an extended time. The modes of input for this multi-person experiment included not only location sensor data for an individual, but also the sensor data on how the individuals interact with each other.

In our Reality Mining project [5], 81 participants wore Nokia 6600 mobile phones with a custom version of the Context software [9] for a period of over nine months, and have their locations (in terms of cell tower usages), proximity information (in terms of the Bluetooth devices seen), cell phone usages (phone calls, short messages, voice mails), and cell phone states continuously collected. From the recorded data set, we are able to infer the participants' social circles, as well as their individual behaviors. As are typical in such large, extended experiments, the cell phone data collection for the individuals experienced several different types of abnormalities, and the cell phone data for different individuals are correlated. Thus this data set provided a good test for our ability to data mine structural relationships among different participants.

The remainder of this article is organized in the following way. In section 2, we formulate the latent structure influence model and give its latent state inference algorithm and its parameter estimation algorithm. In 3 we discuss how the algorithm performs on these example applications, and compare the robustness, accuracy, and efficiency of the method to some other standard approaches to this problem.

2 Influence Modeling of Multi-sensor Dynamics

The influence model is a tractable approximation of the intractable hidden Markov modeling of multiple interacting dynamic processes. While the number of states for the hidden Markov model is the multiplication of the number of states for individual processes, the number of states for the corresponding influence model is the summation of the number of states for individual processes. The influence model attains this tractability by linearly combining the contributions of latent state distributions of individual processes at time t to get the latent state distributions of individual processes at time $t+1$.

In the rest of this section, we describe the influence parameters, the evolution of the marginal latent state distributions for individual processes, and the observations for individual processes as probabilistic functions of the latent states. The usage with an influence model is generally: inference of latent states given parameters and observations, estimation of parameters given latent states and observations, or simultaneous latent state inference and parameter estimation from observations. A graphical model representation of the influence model is plotted in Figure 2.

A latent structure influence process $\{s_t^{(c)}, y_t^{(c)} : c = \{1, \cdots, C\}, t \in \mathbb{N}\}$ is a stochastic process composed of C interacting (sub-) processes. Each process $c \in \{1, \cdots, C\}$ has latent states $\{s_t^{(c)} : t \in \mathbb{N}\}$ and observations $\{y_t^{(c)} : t \in \mathbb{N}\}$ corresponding to sample times $t \in \mathbb{N}$. The latent structure influence process is normally used to estimate the latent states and/or the parameters based on the observations.

The latent state $s_t^{(c)}$ for processes c at time t is a random variable valued over $\{1, \cdots, m_c\}$. The latent state for all C processes at time t is thus $\boldsymbol{s}_t = (s_t^{(1)}, \cdots, s_t^{(C)})$. We write the probability measures of $s_t^{(c)}$ over its values into

a row vector $\boldsymbol{p}(s_t^{(c)})$ and concatenate these row vectors for all processes $c \in \{1, \cdots, C\}$ into a longer row vector $\boldsymbol{p}(\boldsymbol{s}_t)$,

$$\begin{aligned}
\boldsymbol{p}(s_t^{(c)}) &\triangleq \Big(\Pr(s_t^{(c)} = 1), \ldots, \Pr(s_t^{(c)} = m_c)\Big) \\
\boldsymbol{p}(\boldsymbol{s}_t) &\triangleq \Big(\boldsymbol{p}(s_t^{(1)}), \cdots, \boldsymbol{p}(s_t^{(C)})\Big) \\
&= \Big(\Pr(s_t^{(1)} = 1), \ldots, \Pr(s_t^{(1)} = m_1), \cdots, \Pr(s_t^{(C)} = 1), \ldots, \Pr(s_t^{(C)} = m_C)\Big).
\end{aligned}$$

According to the definition of the probability measure, $\sum_{j=1}^{m_c} \Pr(s_t^{(c)} = j) = 1$ for $c \in \{1, \cdots, C\}$.

The probability distributions $P(s_t^{(c)})$ for latent states $s_t^{(c)}$, where $c \in \{1, \cdots, C\}$ and $t \in \{1, \cdots, T\}$, evolve recursively and linearly in a similar way as in the hidden Markov process case. The (initial) probability measure of the latent state $s_{t=1}^{(c)}$ for process $c \in \{1, \cdots, C\}$ at time $t = 1$ over its values is parameterized as a row vector $\pi^{(c)}$, which in turn, is concatenated into a longer row vector $\boldsymbol{\pi}$.

$$\begin{aligned}
\pi_i^{(c)} &\triangleq \Pr(s_1^{(c)} = i), \text{ where } 1 \le i \le m_c \\
\pi^{(c)} &\triangleq \Big(\pi_1^{(c)}, \cdots, \pi_{m_c}^{(c)}\Big) \\
\boldsymbol{\pi} &\triangleq \Big(\pi^{(1)}, \cdots \pi^{(C)}\Big) \\
&= \Big(\pi_1^{(1)}, \cdots \pi_{m_1}^{(1)}, \cdots, \pi_1^{(C)}, \cdots, \pi_{m_C}^{(C)}\Big)
\end{aligned}$$

The probability measures $\boldsymbol{p}(\boldsymbol{s}_t)$ evolve over time linearly as $\boldsymbol{p}(\boldsymbol{s}_{t+1}) = \boldsymbol{p}(\boldsymbol{s}_t) \cdot H$, where H is called an *influence matrix*, as compared to a Markov matrix in a hidden Markov model. The influence matrix H is parameterized in accordance with Asavathiratham's initial parameterization [1]: Call the $C \times C$ matrix $D_{C \times C}$, whose columns each add up to 1, as a *network (influence) matrix*; Call the $m_{c_1} \times m_{c_2}$ Markov matrices $A^{(c_1, c_2)}$ (where $c_1, c_2 \in \{1, \cdots, C\}$), whose rows each add up to 1, as *inter-process state transition matrices*. The influence matrix is formed as the generalized Kronecker product

$$H \triangleq D \otimes \left\{A^{(c_1,c_2)}\right\}_{c_1,c_2 \in \{1,\cdots,C\}} = \left(d_{c_1,c_2} A^{(c_1,c_2)}\right)_{c_1,c_2 \in \{1,\cdots,C\}},$$

which is a block matrix, whose submatrix at row c_1 and column c_2 is $d_{c_1,c_2} A^{(c_1,c_2)}$, and whose element indexed by (c_1, c_2, i, j) is $h_{i,j}^{(c_1,c_2)} = d_{c_1,c_2} \cdot a_{i,j}^{(c_1,c_2)}$. The fact that $\boldsymbol{p}(\boldsymbol{s}_t)$ is a concatenation of probability distributions is guaranteed by the requirement that each column of D, as well as each row of $A^{(c_1,c_2)}$, adds up to 1.

Using this notation, the latent state distributions $\boldsymbol{p}(\boldsymbol{s}_t)$ for the C interacting processes are evolved as

$$\boldsymbol{p}(\boldsymbol{s}_1) = \boldsymbol{\pi}$$
$$\boldsymbol{p}(\boldsymbol{s}_{t+1}) = \boldsymbol{p}(\boldsymbol{s}_t) \cdot H$$

An observation $y_t^{(c)}$ for process c at sample time t is a random variable conditioned on the corresponding latent state $s_t^{(c)}$. The observations are used to adjust the estimation of latent states

$$P(\boldsymbol{s}_t) \cdot P(\boldsymbol{y}_t|\boldsymbol{s}_t) \stackrel{\triangle}{=} \prod_{c=1}^{C} P(s_t^{(c)}) \cdot P(y_t^{(c)}|s_t^{(c)}).$$

When the observation $y_t^{(c)}$ is finite valued, $y_t^{(c)} \in \{1, \ldots, n_c\}$, we write $b_{i,j}^{(c)} = \Pr(y^{(c)} = j|s^{(c)} = i)$ and call the $m_c \times n_c$ matrix $B^{(c)} = (b_{i,j}^{(c)})$ as an *observation matrix*. When the observation $y_t^{(c)}$ is in multivariate normal distribution, we use n_c to represent the dimensionality of $y_t^{(c)}$, and use $\boldsymbol{\mu}_{s^{(c)}}^{(c)}$, $\Sigma_{s^{(c)}}^{(c)}$ to represent the mean and variance of $y_t^{(c)}$. In other words, when the corresponding latent state is valued as $s_t^{(c)}$, we have $y_t^{(c)} \sim N_{n_c}(\mu_{s^{(c)}}^{(c)}, \Sigma_{s^{(c)}}^{(c)})$.

The latent structure influence process $\{s_t^{(c)}, y_t^{(c)} : c = \{1, \cdots, C\}, t \in \mathbb{N}\}$ is a simplification of the hidden Markov process $\{\boldsymbol{s}_t = \left(s_t^{(1)}, \cdots, s_t^{(C)}\right), \boldsymbol{y}_t = \left(y_t^{(1)}, \cdots, y_t^{(C)}\right) : t \in \mathbb{N}\}$. In the hidden Markov process $\{\boldsymbol{s}_t, \boldsymbol{y}_t : t \in \mathbb{N}\}$, a latent state $\boldsymbol{s}_t$ can take $\prod_{c=1}^{C} m_c$ number of values, an observation $\boldsymbol{y}_t$ can observe very complex distributions conditioned on $\boldsymbol{s}_t$, and the state transition matrix is a $(\prod_c m_c) \times (\prod_c m_c)$ Markov matrix. When C is large, the computation on $\{\boldsymbol{s}_t, \boldsymbol{y}_t : t \in \mathbb{N}\}$ becomes intractable, and is easy to overfit. In comparison, in the latent structure influence process, we only need to cope with the marginal probability distributions $\Pr(s_t^{(c)} = i)$ for state $\boldsymbol{s}_t$, and can cope with $y_t^{(c)}$ for individual interacting processes $c \in \{1, \cdots, C\}$ separately. Asavathiratham [1] proved the following theorems concerning the relationship between an influence process and a Markov process: (1) Given any influence process $\{s_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$ parameterized by the initial state distributions $\boldsymbol{\pi} = \boldsymbol{p}(\boldsymbol{s}_1)$ and the influence matrix H, there exists a Markov process $\{\boldsymbol{x}_t = (x_t^{(1)}, \cdots, x_t^{(C)}) : t \in \mathbb{N}\}$ parameterized by the initial state distribution of $\boldsymbol{x}_1$ and the $(\prod_c m_c) \times (\prod_c m_c)$ Markov matrix G, and the corresponding influence process $\{x_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$ has the same probability measure as the original influence process $\{s_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$ (i.e., both influence processes have the same parameters). (2) Given any Markov process $\{\boldsymbol{x}_t = (x_t^{(1)}, \cdots, x_t^{(C)}) : t \in \mathbb{N}\}$ with Markov matrix G, the stochastic process $\{x_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$ is an influence process with influence matrix H. The two matrices are connected by an *event matrix* $B(m_1, \cdots, m_C)$ (where B is determined only by $m_1, \cdots, m_C$), $B \cdot H = G \cdot B$. As a result, the stationary distribution of the Markov process can be linearly mapped into the stationary distribution of the corresponding influence process. We extended Asavathiratham's

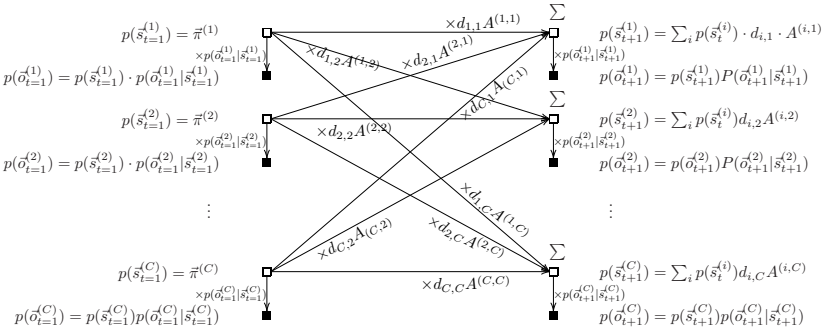


Fig. 2. A graphical model representation of the influence model. The left column represents basis step, and the right column represents the induction step. Shadowed squares are observable, while un-shadowed squares are unobservable. Our task is to learn the parameters and latent states from observations. The two-column convention is adopted from Murphy [6].

influence process $\{s_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$ into a latent structure influence process $\{s_t^{(c)}, y_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$, and use the latent structure influence process to understand/simulate how a group of experts cooperate with each other and make predictions.

The forward-backward algorithm for latent state estimation and the maximum likelihood algorithm for parameter estimation for an influence model are derived from the equivalence of the influence model and the corresponding hidden Markov model. Being able to model the dynamics of C interacting processes, with m_c number of latent states for individual process c, in a polynomial complexity in the sum of the number of latent states for individual chains $O(\sum m_c)$ does not necessarily guarantee that the latent state estimation and the parameter estimation algorithms also have a polynomial time complexity. We give the latent state estimation (E-step), as well as the parameter estimation algorithm (M-step) in Algorithm 1. The derivation is deferred in the Appendix. In this algorithm, the random variables $y_t^{(c)}$ for $c \in \{1, \cdots, C\}$ and $t \in \{1, \cdots, T\}$ are already sampled and their values are known. We write the probability (or probability density) of observing $y_t^{(c)}$ when the latent state $s_t^{(c)}$ take values $1, \cdots, m_c$ into a $m_c \times 1$ row vector $\boldsymbol{b}_t^{(c)} \triangleq \left(\Pr(y_t^{(c)}|s_t^{(c)} = 1), \cdots, \Pr(y_t^{(c)}|s_t^{(c)} = m_c)\right)$, and concatenate them into a $(\sum_c m_c) \times 1$ row vector $\boldsymbol{b}_t$. The quantities $\alpha_{t,i}^{(c)} \triangleq \Pr(s_t^{(c)} = i | \{y_{t_0}^{(c_0)} : c_0 \in \{1, \cdots, C\}, t_0 \in \{1, \cdots, t\}\})$ are *forward parameters*. We write $\alpha_{t,i}^{(c)}$ for all $i \in \{1, \cdots, m_c\}$ into a $m_c \times 1$ row vector $\boldsymbol{\alpha}_t^{(c)}$, and concatenate $\boldsymbol{\alpha}_t^{(c)}$ into a $(\sum_c m_c) \times 1$ row vector $\boldsymbol{\alpha}_t$. The quantities $\beta_{t,i}^{(c)} \triangleq \Pr(\{y_{t_0}^{(c_0)} : c_0 \in \{1, \cdots, C\}, t_0 \in \{t+1, \cdots, T\}\} | s_t^{(c)} = i)$ are *backward parameters*. We write $\beta_{t,i}^{(c)}$ for all $i \in \{1, \cdots, m_c\}$ into a $1 \times m_c$ column vector

$\boldsymbol{\beta}_t^{(c)}$, and concatenate $\boldsymbol{\beta}_t^{(c)}$ into a $1 \times (\sum_c m_c)$ column vector $\boldsymbol{\beta}_t$. The quantities $\gamma_{t,i}^{(c)} \triangleq \Pr(s_t^{(c)} = i | \{y_{t_0}^{(c_0)} : c_0 \in \{1, \cdots, C\}, t_0 \in \{1, \cdots, T\}\})$ are *one-slice parameters*. We write $\gamma_{t,i}^{(c)}$ for all $i \in \{1, \cdots, m_c\}$ into a $m_c \times 1$ row vector $\boldsymbol{\gamma}_t^{(c)}$, and concatenate $\boldsymbol{\gamma}_t^{(c)}$ into a $(\sum_c m_c) \times 1$ row vector $\boldsymbol{\gamma}_t$. The quantities $\xi_{t-1 \to t, i \to j}^{(c_1,c_2)} \triangleq \Pr(s_{t-1}^{(c_1)} = i, s_t^{(c_2)} = j | \{y_{t_0}^{(c_0)} : c_0 \in \{1, \cdots, C\}, t_0 \in \{1, \cdots, T\}\})$ are *two-slice parameters*. We write $\xi_{t-1 \to t, i \to j}^{(c_1,c_2)}$ for all $i \in \{1, \cdots, m_{c_1}\}$ and $j \in \{1, \cdots, m_{c_2}\}$ into a $m_{c_1} \times m_{c_2}$ matrix $\xi_{t-1 \to t}^{(c_1,c_2)}$, and concatenate $\xi_{t-1 \to t}^{(c_1,c_2)}$ into a $(\sum_c m_c) \times (\sum_c m_c)$ matrix whose submatrix at row c_1 and column c_2 is $\xi_{t-1 \to t}^{(c_1,c_2)}$.

3 Experimental Results

In this section, we illustrate how an influence model can capture the correlations among different dynamic classification processes. We will show how capturing the correct structure between different "experts" can allow improvement of the overall classification performance. We will also illustrate the efficiency and robustness to noise that this modeling capability provides.

We begin with a synthetic example of a noisy sensor net in order to illustrate the structure that the influence model tries to capture, and how an influence model can be used to improve classification precision. We then extend the noisy body sensor net example and compare the training errors and the testing errors of different dynamic models.

We will then show application of the algorithm to two real examples:

The first example is a wearable smart sensor net example in which the goal is real-time context recognition, and the influence model is used to discover hidden structure among speech, location, activity, and posture classification experts in order to allow for more accurate and robust classification of the wearer's overall state.

The second example is a group of 81 people carrying smart phones that are programmed to record location, proximity to other experimental subjects, and cell phone usage. In this example we will focus on the ability of the influence model to correctly determine the social structure of the group.

3.1 Combining Evidence with the Influence Model

Central to the latent structure influence model is the mechanism that the evidence is combined in the latent state level over time. This mechanism both enables coping with hetereogeneous types of evidence, and makes it possible to automatically find out relations among the different pieces of evidence.

In this subsection, we use a simple example involving two Gaussian distributions to illustrate how the influence model combine evidence and set priors for individual related processes ("experts"). We also compare the mechanisms of the latent structure influence model, the hidden Markov model with full covariant matrix, and the hidden Markov model with diagonal covariant matrix.

Algorithm 1. The EM algorithm for the latent structure influence model

E-Step

$$\boldsymbol{\alpha}_t^* = \begin{cases} \boldsymbol{\pi}_{1\times\sum m_c} \cdot \mathtt{diag}[\boldsymbol{b}_1] & t=1 \\ \boldsymbol{\alpha}_{t-1} \cdot H \cdot \mathtt{diag}[\boldsymbol{b}_t] & t>1 \end{cases}$$

$$\mathcal{N}_t = \mathtt{diag}\left[\left(\underbrace{\frac{1}{\sum_{i=1}^{m_1}\boldsymbol{\alpha}_{t,i}^{*(1)}},\cdots,\frac{1}{\sum_{i=1}^{m_1}\boldsymbol{\alpha}_{t,i}^{*(1)}}}_{m_1},\cdots,\underbrace{\frac{1}{\sum_{i=1}^{m_C}\boldsymbol{\alpha}_{t,i}^{*(C)}},\cdots,\frac{1}{\sum_{i=1}^{m_C}\boldsymbol{\alpha}_{t,i}^{*(C)}}}_{m_C}\right)\right]$$

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t^* \cdot \mathcal{N}_t$$

$$\boldsymbol{\beta}_t = \begin{cases} \mathbf{1}_{\sum m_c\times 1} & t=T \\ H \cdot \mathtt{diag}[\boldsymbol{b}_t] \cdot \mathcal{N}_{t+1} \cdot \boldsymbol{\beta}_{t+1} & t<T \end{cases}$$

$$\boldsymbol{\gamma}_t = \boldsymbol{\alpha}_t \cdot \mathtt{diag}[\boldsymbol{\beta}_t]$$

$$\xi_{t-1\to t} = \mathtt{diag}[\boldsymbol{\alpha}_{t-1}] \cdot H \cdot \mathtt{diag}[\boldsymbol{b}_t] \cdot \mathcal{N}_t \cdot \mathtt{diag}[\boldsymbol{\beta}_t]$$

$$p(\boldsymbol{y}) = \prod_{t,c}(\sum_{i=1}^{m_c}\boldsymbol{\alpha}_{t,i}^{*(c)})$$

M-Step

– Parameters related to the latent state transitions

$$A^{(i,j)} = \mathtt{normalize}[\sum_{t=2}^{T}\xi_{t-1\to t}^{(i,j)}]$$

$$S = \begin{pmatrix} \mathbf{1}_{1\times m_1} & & \\ & \cdots & \\ & & \mathbf{1}_{1\times m_C} \end{pmatrix}$$

$$d_{ij} = \mathtt{normalize}[S\left(\sum_{t=2}^{T}\xi_{t-1\to t}\right)S^{\mathrm{T}}]$$

$$\boldsymbol{\pi}^{(c)} = \mathtt{normalize}[\boldsymbol{\gamma}_1^{(c)}]$$

– Parameters related to multinomial observations

$$B^{(c)} = \mathtt{normalize}[\sum_t \boldsymbol{\gamma}_t^{(c)\,\mathrm{T}} \cdot \left(\delta(y_t^{(c)},1),\cdots,\delta(y_t^{(c)},\cdots,n_c)\right)]$$

– Parameters related to Gaussian observations

$$\boldsymbol{\mu}^{(c)} = \left(\sum_t \boldsymbol{\gamma}_t^{(c)\,\mathrm{T}} \cdot \boldsymbol{y}_t^{(c)}\right) / \left(\sum_t \boldsymbol{\gamma}_t^{(c)} \cdot \mathbf{1}_{m_c\times 1}\right)$$

$$\Sigma_i^{(c)} = \left(\sum_t \gamma_{t,i}^{(c)} \boldsymbol{y}_t^{(c)} \boldsymbol{y}_t^{(c)\,\mathrm{T}}\right) / \left(\sum_t \boldsymbol{\gamma}_t^{(c)} \cdot \mathbf{1}_{m_c\times 1}\right) - \boldsymbol{\mu}_i^{(c)} \cdot \boldsymbol{\mu}_i^{(c)\,\mathrm{T}}$$

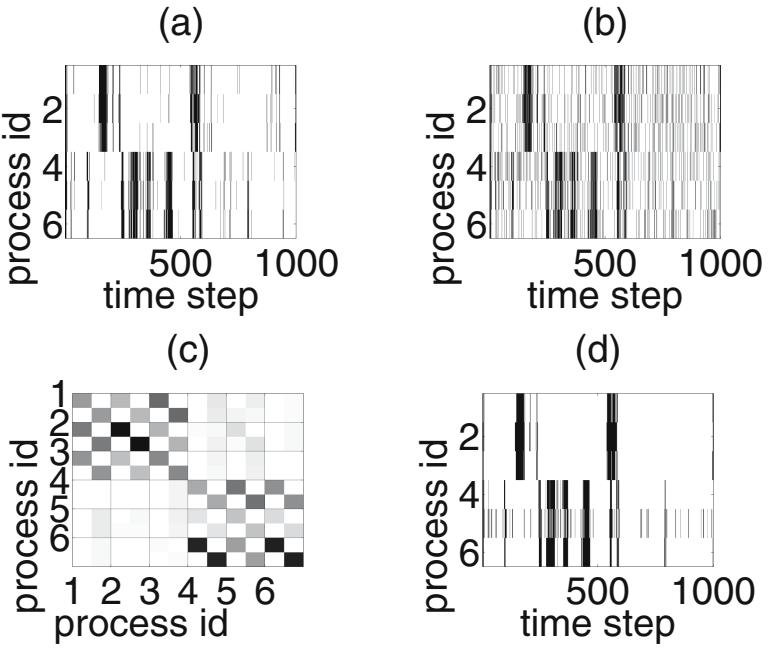


Fig. 3. Inference from observations of interacting dynamic processes

Noisy Body Sensor Net Example. In the noisy body sensor net example, we have six stochastic processes, and we sample these six processes with six body sensors. Each process can be either signaled (one) or non-signaled (zero) at any time, and the corresponding body sensor has approximately 10% of its samples flipped. The interaction of the six stochastic processes behind the scene looks like this: processes one through three tend to have the same states; processes four through six tend to have the same states; the processes are more likely to be non-signaled than to be signaled; and the processes tend to stick to their states for a stretch of time. The parameters of the model are given as the following and are going to be estimated: $A_{ij} = \begin{pmatrix} .99 & .01 \\ .08 & .92 \end{pmatrix}, 1 \leq i, j \leq 6$, $B_i = \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}, 1 \leq i \leq 6$, $d_{ij} = .33, 1 \leq i, j \leq 3$, and $d_{ij} = .33, 4 \leq i, j \leq 6$.

In Figure 3, (a) shows the sampled latent state sequences, (b) shows the corresponding observation sequences, (c) shows the influence matrix reconstructed from sampled observation sequences, and (d) shows the reconstructed latent state sequences after 300 observations. The $(i, j)^{th}$ entry of the $(c_1, c_2)^{th}$ sub-matrix of an influence matrix determines how likely that process c_1 is in state i at time t and process c_2 is in state j at time $t + 1$. It can be seen from Figure 3 (c) that the influence model computation recovers the structure of the interaction.

The influence model can normally attain around 95% accuracy in predicting the latent states for each process. The reconstructed influence matrix has only

9% relative differences with the original one. Using only observations of other chains we can predict a missing chain's state with 87% accuracy.

Comparison of Dynamic Models. The training errors and the testing errors of the coupled hidden Markov model, the hidden Markov model, and the influence model are compared in this example. The setup of the comparison is described as the following. We have a Markov process with 2^C, where $C = 10$, number of states and a randomly generated state transition matrix. Each system state $\boldsymbol{s}_t$ is encoded into a binary $s_t^{(1)} \cdots s_t^{(C)}$. Each of the $m_c = 2$ evaluations of "digit" $s_t^{(c)}$corresponds a different 1-d Gaussian observation $o_t^{(c)}$: Digit $s_t^{(c)} = 1$ corresponds to $o_t^{(c)} \sim \mathcal{N}[\mu_1 = 0, \sigma_1^2 = 1]$; Digit $s_t^{(c)} = 2$ corresponds to $o_t^{(c)} \sim \mathcal{N}[\mu_2 = 1, \sigma_2^2 = 1]$.

In most real sensor nets we normally have redundant measures and an insufficient observations to accurately characterize sensor redundancy using standard methods. Figure 4 compares the performances of several dynamic latent structure models applicable to multi-sensor systems. Of the 1000 samples $(\boldsymbol{o}_t)_{1 \le t \le 100}$, we use the first 250 for training and all 1000 for validation.

There are two interesting points. First, the logarithmically scaled number of parameters of the influence model allows us to attain high accuracy based on a relatively small number of observations. This is because the eigenvectors of the master Markov model we want to approximate are either mapped to the eigenvectors of the corresponding influence model, or mapped to the null space of the corresponding event matrix thus is not observable from the influence model, and

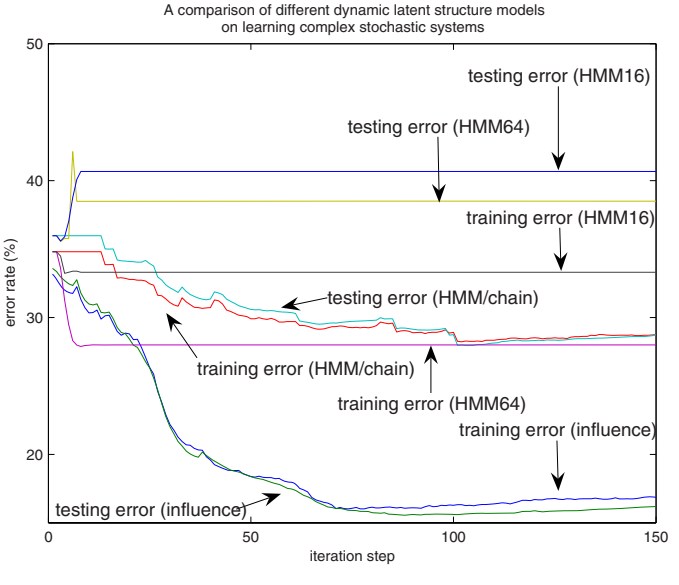


Fig. 4. Comparison of dynamic models

that in addition the eigenvector with the largest eigenvalue (i.e., 1) is mapped to the eigenvector with the largest eigenvalue of the influence matrix [1].

Secondly, both the influence model and the hidden Markov model applied to individual processes are relatively immune to over-fitting, at the cost of low convergence rates. This situation is intuitively the same as the numerical analysis wisdom that a faster algorithm is more likely to converge to a local extremum or to diverge.

3.2 On-Body Smart Sensor Network

In the *inSense* system the sensors consist of a chest-mounted camera, a Wi-Fi transceiver, an ambient audio recorder, and two accelerometers, worn on hip and wrist (see Figure 1) [3]. This system is designed to classify in real time eight locations, six speaking/non-speaking status, seven postures, and eight activities. The classification is carried out in two steps: A pre-classifier (either a single Gaussian, mixture of Gaussians, or C4.5 classifier) is first invoked on the audio and accelerometer features to get a moderately accurate pre-classification result within each the above four categories. These are the "experts" that we desire to group in order to produce more accurate estimates of the wearer's context.

The pre-classification result of different categories is then fed into an influence model to learn inter-sensor structure, and then this learned structure is used to generate an improved post-classification result. In this example the influence model learns the conditional probabilities that relate the four categories (location, audio, posture, and activity) and then uses this learned influence matrix to improve the overall performance.

For example, given that the *inSense* wearer is typing, we can inspect the row of the influence matrix corresponding to "typing" and see that this person is very likely to be either in the office or at home, to not be speaking, and to be sitting. As a result, the action of typing can play a critical role to disambiguating confusions between sitting and standing, or between speaking vs not-speaking, but not between office and home.

By combining evidence across different categories using the influence model, the classification errors for locations, speaking/non-speaking, postures, and activities decreased by an average of 23%, from 38%, 22%, 8% and 27% to 28%, 19%, 8%, and 17% respectively. The post-classification for postures does not show significant improvement because of two reasons: (1) it is already precise enough considering that we have labeling imprecision in our training data and testing data, and (2) it is the driving force for improving the other categories, and no other categories are more certain than the posture category.

3.3 Social Network Example

This example demonstates reconstructing the social structure of a set of subjects from their cellphone-collected data [5]. In this data 81 subjects wore Bluetooth-enabled mobile telephones that recorded which cell towers were visible to the telephone, thus allowing coarse estimation of the wearers' location, and which

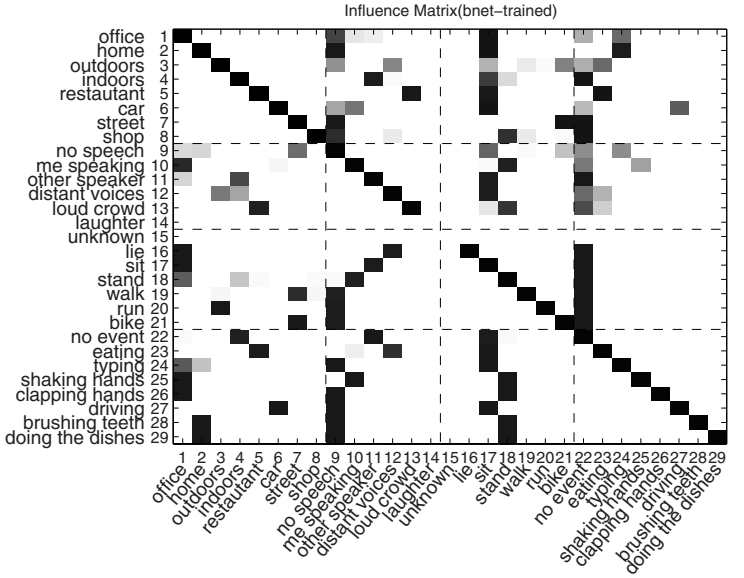


Fig. 5. Influence matrix learned by the EM algorithm

Bluetooth devices are nearby, thus allowing inference of proximity to other subjects. Note that Bluetooth signals include a unique identifier, and are typically detectable at a range of only a few meters. In this study fixed Bluetooth beacons were also employed, allowing fairly precise estimation of subjects location even within buildings. Over the nine months of the study 500,000 hours of data were recorded.

The temporal evolution of these observations were analyzed using the influence model with 81 chains, corresponding to the 81 subjects. Each subjects' chain was constrained to have two latent states ("work", "home") but with no restriction on social network connectivity.

In our first experiment with this data the observation vector for each chain was restricted to the cell tower visibility of each subjects' 10 most commonly seen cell towers. In the resulting model the two states for each subject corresponded accurately to 'in the office' and 'at home', with other locations being misclassified. The resulting influence matrix, shown in Figure 6 (a), demonstrated that most people follow very regular patterns of movement and interpersonal association, and consequently we can predict their actions with substantial accuracy from observations of the other subjects. A few of the chains were highly independent and thus not predictable. These chains corresponded to new students, who had not yet picked up the rhythm of the community, and the faculty advisors, whose patterns are shown to determine the patterns of other students.

In another setup, we used the Bluetooth proximity distribution as our observations. Again, the latent states accurately reflect whether a person is at home

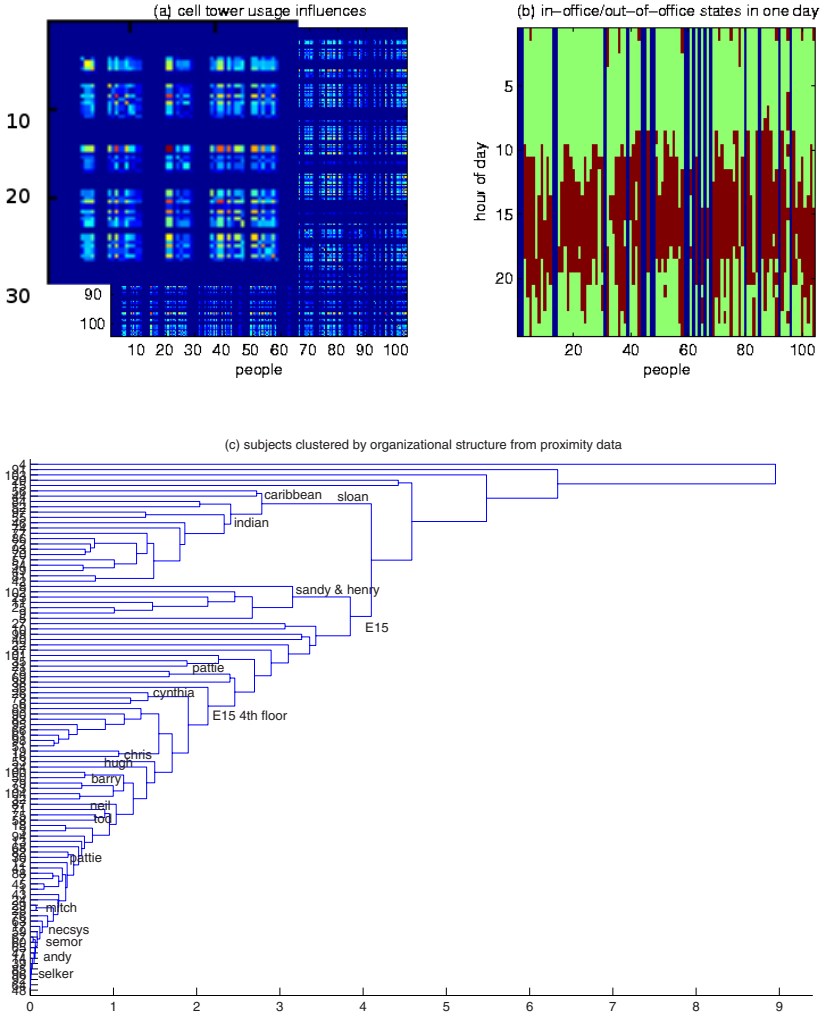


Fig. 6. Finding social structures from cellphone-collected data. (a) New students and faculty are outliers in the influence matrix, appearing as red dots due to large self-influence values. (b) Most People follow regular patterns (red: in office, green: out of office, blue: no data), (c) clustering influence values recovers workgroup affiliation with high accuracy (labels show name of group).

of in office. However with this data the resulting influence matrix shows precisely the social and geometrical structure of the subjects. The dendrogram from the proximity influence matrix shown in Figure 6 (b) captures the actual organization of the laboratory, clustering people into their actual work groups with only three errors. For comparison, a clustering based on direct correlations in the data has six errors.

4 Conclusion

We have presented the formulation of a latent structure influence model, given the parameter learning and latent state estimation algorithms, and demonstrated the latent structure influence model's performance in combining and analyzing networks of experts. Both in simulation and real examples the influence model proved to be an efficient and accurate method of estimating unknown network structure and simultaneously estimating transition parameters. This was shown to allow more accurate estimates of state, and increased tolerance to missing data and similar noise. As a result, we believe that the latent structure influence process will provide a good framework for human computing applications.

Matlab code for the influence model and for the synthetic sensor net example may be found at:
http://vismod.media.mit.edu/vismod/demos/influence-model/index.html.

References

[1] Chalee Asavathiratham. *The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains*. PhD thesis, MIT, 1996.
[2] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, and Alex Pentland. Learning human interactions with the influence model. Technical report, MIT Media Laboratory Vision & Modeling Technical Report #539, 2001. URL http://vismod.media.mit.edu/tech-reports/TR-539.pdf.
[3] Mark Blum, Alex Pentland, and Gehrard Tröster. Insense: Interest-based life logging. In *IEEE Multimedia*, volume 13(4), pages 40–48, 2006.
[4] DARPA. Assist proposer information pamphlet, 2004. URL http://www.darpa.mil/ipto/solicitations/open/04-38_PIP.htm.
[5] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous Computing*, 2005.
[6] Kevin Murphy. The bayes net toolbox for matlab. In *Computing Science and Statistics*, 2001.
[7] Nuria M. Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interactions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22(8), pages 831–843, 2000.
[8] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas Huang. Human computing and machine understanding of human behavior: A survey. In *Proceedings of the 8th International Converence on Multimodal Interfaces*, pages 239–248, 2006.
[9] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. Contextphone — a prototyping platform for context-aware mobile applications. In *IEEE Pervasive Computer*, April 2005.

Appendix: A Derivation of the EM Algorithm for the Influence Process

In section 2, we gave the EM algorithm for the latent structure influence process. The derivation of this algorithm is below.

Latent State Inference

The task of latent state inference is to estimate

$$\left(s_t^{(c)}\right)_{1\le t\le T}^{1\le c\le C} \triangleq \left\{s_t^{(c)} : c \in \{1,\cdots,C\}, t \in \{1,\cdots,T\}\right\}$$

from

$$\left(y_t^{(c)}\right)_{t+1\le t\le T}^{1\le c\le C} \triangleq \left\{y_t^{(c)} : c \in \{1,\cdots,C\}, t \in \{1,\cdots,T\}\right\}$$

given the influence parameters.

Theorem 1. *Let the marginal forward parameters $\alpha_t^{(c)}(s_t^{(c)})$, the marginal backward parameters $\beta_t^{(c)}(s_t^{(c)})$, the marginal one-slice parameters $\gamma_t^{(c)}(s_t^{(c)})$, the marginal two-slice parameters $\xi_{t\to t+1}^{(c_1,c_2)}(s_t^{(c_1)}, s_{t+1}^{(c_2)})$ of a latent structure influence model be*

$$\begin{aligned}
\alpha_t^{(c)}(s_t^{(c)}) &= P\left(s_t^{(c)}, \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le t}^{1\le c_1\le C}\right)\\
\beta_t^{(c)}(s_t^{(c)}) &= P\left(\left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \Big| s_t^{(c)}\right)\\
\gamma_t^{(c)}(s_t^{(c)}) &= P\left(s_t^{(c)} \Big| \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le T}^{1\le c_1\le C}\right)\\
\xi_{t\to t+1}^{(c_1,c_2)}(s_t^{(c_1)}, s_{t+1}^{(c_2)}) &= P\left(s_t^{(c_1)} s_{t+1}^{(c_2)} \Big| \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le T}^{1\le c_1\le C}\right)
\end{aligned}$$

They can be computed recursively in the following way:

$$\begin{aligned}
\alpha_1^{(c)}(s_1^{(c)}) &= P\left(\left(y_t^{(c)}\right)^{1\le c\le C} | s_1^{(c)}\right)\cdot \pi_{s_1^{(c)}}^{(c)}\\
\alpha_t^{(c)}(s_{2\le t}^{(c)}) &= P\left(\left(y_t^{(c)}\right)^{1\le c\le C} | s_t^{(c)}\right) \sum_{c_1, s_{t-1}^{(c_1)}} \alpha(s_{t-1}^{(c_1)}) h_{s_{t-1}^{(c_1)} s_t^{(c)}}^{(c_1,c)}\\
\beta_T^{(c)}(s_T^{(c)}) &= 1\\
\beta_{t<T}^{(c)}(s_t^{(c)}) &= \frac{1}{C}\cdot \sum_{c_1=1}^{C}\sum_{s_{t+1}^{(c_1)}=1}^{m_{c_1}} h_{s_t^{(c)}, s_{t+1}^{(c_1)}}^{(c,c_1)} \cdot P\left(\left(y_{t+1}^{(c)}\right)^{1\le c\le C} | s_{t+1}^{(c_1)}\right)\beta_{t+1}^{(c)}(s_{t+1}^{(c)})\\
\gamma_t^{(c)}(s_t^{(c)}) &= \alpha_t^{(c)}(s_t^{(c)})\cdot \beta_t^{(c)}(s_t^{(c)})\\
\xi_{t\to t+1}^{(c_1,c_2)}(s_t^{(c_1)}, s_{t+1}^{(c_2)}) &= \alpha_t^{(c_1)}(s_t^{(c_1)})\cdot h_{s_t^{(c_1)}, s_{t+1}^{(c_2)}}^{(c_1,c_2)} \cdot \beta_{t+1}^{(c_2)}(s_{t+1}^{(c_2)}) \cdot P\left(\left(y_{t+1}^{(c)}\right)^{1\le c\le C} | s_{t+1}^{(c_2)}\right)
\end{aligned}$$

Proof. In the following, we demonstrate that we can solve for the marginal forward parameters without first solving the joint marginal forward parameters.

– Basis Step

$$\begin{aligned}
&\alpha(s_1^{(c)}) \\
&= P\left(s_t^{(c)}, \left(y_1^{(c_1)}\right)^{1\le c_1\le C}\right) \\
&= P\left(\left(y_1^{(c_1)}\right)^{1\le c_1\le C} | s_1^{(c)}\right) \cdot P\left(s_1^{(c)}\right) \\
&= P\left(\left(y_1^{(c_1)}\right)^{1\le c_1\le C} | s_1^{(c)}\right) \cdot \pi^{(c)}_{s_1^{(c)}}
\end{aligned}$$

– Induction Step

$$\begin{aligned}
&\alpha(s_{t\ge 2}^{(c)}) \\
&= P\left(s_t^{(c)}, \left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le t}\right) \\
&= P\left(\left(y_t^{(c_1)}\right)^{1\le c_1\le C} | s_t^{(c)}\right) \cdot P\left(s_t^{(c)}, \left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le t-1}\right) \\
&= P\left(\left(y_t^{(c_1)}\right)^{1\le c_1\le C} | s_t^{(c)}\right) \cdot \sum_{c_1=1}^{C}\sum_{s_{t-1}^{(c_1)}=1}^{m_{c_1}} P\left(s_{t-1}^{(c_1)}, \left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le t-1}\right) \cdot h^{(c_1,c)}_{s_{t-1}^{(c_1)} s_t^{(c)}} \\
&= P\left(\left(y_t^{(c_1)}\right)^{1\le c_1\le C} | s_t^{(c)}\right) \cdot \left(\sum_{c_1=1}^{C}\sum_{s_{t-1}^{(c_1)}=1}^{m_{c_1}} \alpha(s_{t-1}^{(c_1)}) \cdot h^{(c_1,c)}_{s_{t-1}^{(c_1)} s_t^{(c)}}\right)
\end{aligned}$$

In the following, we show that we can get the marginal backward parameters without the knowledge of the joint backward parameters.

– Basis Step. We have $\beta(s_T^{(c)}) = 1$ trivially, and

$$\begin{aligned}
\sum_{s_T^{(c)}=1}^{m_c} \alpha(s_T^{(c)}) \cdot \beta(s_T^{(c)}) &= \sum_{s_T^{(c)}=1}^{m_c} P\left(s_T^{(c)}, \left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le T}\right) \\
&= P\left(\left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le T}\right) \\
\frac{1}{C} \cdot \sum_{c=1}^{C}\sum_{s_T^{(c)}=1}^{m_c} \alpha(s_T^{(c)}) \cdot \beta(s_T^{(c)}) &= P\left(\left(y_{t_1}^{(c_1)}\right)^{1\le c_1\le C}_{1\le t_1\le T}\right)
\end{aligned}$$

– Induction Step

$$
\begin{aligned}
&\beta(s_{t<T}^{(c)}) \\
&= P\left(\left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \middle| s_t^{(c)}\right) \\
&= \sum_{s_{t+1}^{(c_1)}=1}^{m_C} P\left(s_{t+1}^{(c_1)}, \left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \middle| s_t^{(c)}\right), 1 \le c_1 \le C \\
&= \frac{1}{C}\cdot \sum_{c_1=1}^{C}\sum_{s_{t+1}^{(c_1)}=1}^{m_C} P\left(s_{t+1}^{(c_1)}, \left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \middle| s_t^{(c)}\right) \\
&= \frac{1}{C}\cdot \sum_{c_1=1}^{C}\sum_{s_{t+1}^{(c_1)}=1}^{m_C} P\left(\left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \middle| s_{t+1}^{(c_1)}\right)\cdot P(s_{t+1}^{(c_1)}|s_t^{(c)}) \\
&= \frac{1}{C}\cdot \sum_{c_1=1}^{C}\sum_{s_{t+1}^{(c_1)}=1}^{m_{c_1}} \beta(s_{t+1}^{(c_1)})\cdot h_{s_t^{(c)} s_{t+1}^{(c_1)}}^{(c_1,c)} \cdot P\left(\left(y_t^{(c_1)}\right)^{1\le c_1\le C} | s_{t+1}^{(c_1)}\right)
\end{aligned}
$$

The one-slice parameters $\gamma_t^{(c)}(s_t^{(c)})$ can be computed from the marginal forward parameters and the marginal backward parameters

$$
\begin{aligned}
\gamma_t^{(c)}(s_t^{(c)}) &= P\left(s_t^{(c)}, \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le T}^{1\le c_1\le C}\right) \\
&= P\left(s_t^{(c)}, \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le t}^{1\le c_1\le C}\right) P\left(\left(y_{t_1}^{(c_1)}\right)_{t+1\le t_1\le T}^{1\le c_1\le C} \middle| s_t^{(c)}\right) \\
&= \alpha_t^{(c)}(s_t^{(c)})\cdot \beta_t^{(c)}(s_t^{(c)})
\end{aligned}
$$

The two-slice parameters $\xi_{t\to t+1}^{(c_1,c_2)}(s_t^{(c_1)}, s_{t+1}^{(c_2)})$ can also be computed from the marginal forward parameters $\alpha_t^{(c)}(s_t^{(c)})$ and the marginal backward parameters $\beta_t^{(c)}(s_t^{(c)})$:

$$
\begin{aligned}
\xi_{t\to t+1}^{(c_1,c_2)}(s_t^{(c_1)}, s_{t+1}^{(c_2)}) &= P\left(s_t^{(c_1)} s_{t+1}^{(c_2)}, \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le T}^{1\le c_1\le C}\right) \\
&= P\left(s_t^{(c_1)} \left(y_{t_1}^{(c_1)}\right)_{1\le t_1\le t}^{1\le c_1\le C}\right)\cdot P\left(s_{t+1}^{(c_2} | s_t^{(c_1)}\right)\cdot \\
&\quad P\left(\left(y_{t+1}^{(c_1)}\right)^{1\le c_1\le C} \middle| s_{t+1}^{(c_2)}\right)\cdot P\left(\left(y_{t_1}^{(c_1)}\right)_{t+2\le t_1\le T}^{1\le c_1\le C} \middle| s_{t+1}^{(c_2)}\right) \\
&= \alpha_t^{(c_1)}\cdot h_{s_t^{(c_1)} s_{t+1}^{(c_2)}}^{(c_1,c_2)} \cdot P\left(\left(y_{t+1}^{(c_1)}\right)^{1\le c_1\le C} \middle| s_{t+1}^{(c_2)}\right)\cdot \beta_{t+1}^{(c_2)}
\end{aligned}
$$

Parameter Estimation

Suppose the latent states at time $t = 1..T$ is already known $\boldsymbol{s}_t = s_t^{(1)} \cdots s_t^{(C)}$. The likelihood function is

$$
\begin{aligned}
& P\left(\left(y_t^{(c)}\right)_{1\le t\le T}^{1\le c\le C}\right) \\
&= \pi_{\boldsymbol{S}_1} \cdot \left(\prod_{t=1}^{T-1} g_{\boldsymbol{s}_t \to \boldsymbol{s}_{t+1}}\right) \cdot \left(\prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{s}_t)\right) \\
&= \left(\prod_{c=1}^{C} \pi^{(c)}_{s_1^{(c)}}\right) \cdot \left(\prod_{t=1}^{T}\prod_{c2=1}^{C}\sum_{c1=1}^{C} h^{(c1,c2)}_{s_t^{(c1)},s_{t+1}^{(c2)}}\right) \cdot \left(\prod_{t=1}^{T}\prod_{c=1}^{C} P(y_t^{(c)}|s_t^{(c)})\right)
\end{aligned}
$$

We can find new parameters and try to maximize the log likelihood function:

$$
\begin{aligned}
& \log P\left(\left(y_t^{(c)}\right)_{1\le t\le T}^{1\le c\le C}\right) \\
&= \sum_{c=1}^{C} \log \pi^{(c)}_{s_1^{(c)}} + \sum_{t=1}^{T}\sum_{c=1}^{C} \log P(y_t^{(c)}|s_t^{(c)}) + \sum_{t=1}^{T}\sum_{c2=1}^{C} \log \sum_{c1=1}^{C} h^{(c1,c2)}_{s_t^{(c1)},s_{t+1}^{(c2)}} \\
&\ge \sum_{c=1}^{C} \log \pi^{(c)}_{s_1^{(c)}} + \sum_{t=1}^{T}\sum_{c=1}^{C} \log P(y_t^{(c)}|s_t^{(c)}) + \sum_{t=1}^{T}\sum_{c2=1}^{C}\sum_{c1=1}^{C} \log h^{(c1,c2)}_{s_t^{(c1)},s_{t+1}^{(c2)}} \qquad (1) \\
&= \sum_{c=1}^{C}\sum_{i=1}^{m_c} \delta(s_1^{(c)}, i) \cdot \log \pi_i^{(c)} + \sum_{t=1}^{T}\sum_{c=1}^{C}\sum_{i=1}^{m_c} \delta(s_t^{(c)}, i) \cdot \log P(y_t^{(c)}|i) + \qquad (2) \\
&\quad \sum_{t=1}^{T-1}\sum_{c=1}^{C}\sum_{c1=1}^{C}\sum_{i=1}^{m_{c1}}\sum_{j=1}^{m_{c2}} \delta(s_t^{(c1)}, i) \cdot \delta(s_{t+1}^{(c2)}, j) \cdot \log h_{i,j}^{(c1,c2)} \\
&\triangleq \sum_{c=1}^{C}\sum_{i=1}^{m_c} \tilde{\pi}_i^{(c)} \cdot \log \pi_i^{(c)} + \sum_{c=1}^{C}\sum_{i=1}^{m_c} \tilde{\gamma}^{(c)}(i) \cdot \log P(y_t^{(c)}|i) + \\
&\quad \sum_{c=1}^{C}\sum_{c1=1}^{C}\sum_{i=1}^{m_{c1}}\sum_{j=1}^{m_{c2}} \tilde{\xi}^{(c1,c2)}(i,j) \cdot \log h_{i,j}^{(c1,c2)}
\end{aligned}
$$

where the step 1 is according to the Jensen's inequality, and the function $\delta(i,j) = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases}$ is the Kronecker delta function. From 2, we know that $\tilde{\pi}_i^{(c)} = \delta(s_1^{(c)}, i)$, $\tilde{\xi}^{(c1,c2)}(i,j) = \sum_{t=1}^{T-1} \delta(s_t^{(c1)}, i) \cdot \delta(s_{t+1}^{(c2)}, j)$, and $\tilde{\gamma}^{(c)}(i) = \sum_{t=1}^{T} \delta(s_t^{(c)}, i)$ are the sufficient statistics for $\pi_i^{(c)}$, $h_{i,j}^{(c1,c2)}$, and $P(y_t^{(c)}|i)$ respectively. We can maximize the parameters involved in the influence matrix H by equaling them to the corresponding sufficient statistics:

$$\pi_i^{(c)} = \tilde{\pi}_i^{(c)} \qquad (3)$$

$$h_{i,j}^{(c1,c2)} = \frac{1}{C} \cdot \frac{\tilde{\xi}_{i,j}^{(c1,c2)}}{\sum_{j=1}^{m_{c2}} \tilde{\xi}_{i,j}^{(c1,c2)}} \qquad (4)$$

We can maximize the parameters that map the latent states to the observations in the same way as in an ordinary hidden Markov model.

When the latent states at time $t = 1..T$ are not known. We can choose parameters that maximize the expected log likelihood function:

$$
\begin{aligned}
&\mathtt{E}_{\boldsymbol{s}_1 \cdots \boldsymbol{s}_T}\left[\log p\left(\left(y_t^{(c)}\right)_{1\le t\le T}^{1\le c\le C}\right)\right] \\
&= \mathtt{E}_{\boldsymbol{s}_1 \cdots \boldsymbol{s}_T}\left[\sum_{c=1}^{C} \log \pi^{(c)}_{s_1^{(c)}} + \sum_{t=1}^{T}\sum_{c2=1}^{C} \log \sum_{c1=1}^{C} h^{(c1,c2)}_{s_t^{(c1)}, s_{t+1}^{(c2)}} + \sum_{t=1}^{T}\sum_{c=1}^{C} \log P(y_t^{(c)}|s_t^{(c)})\right] \\
&\ge \mathtt{E}_{\boldsymbol{s}_1 \cdots \boldsymbol{s}_T}\left[\sum_{c=1}^{C} \log \pi^{(c)}_{s_1^{(c)}} + \sum_{t=1}^{T}\sum_{c2=1}^{C}\sum_{c1=1}^{C} \log h^{(c1,c2)}_{s_t^{(c1)}, s_{t+1}^{(c2)}} + \sum_{t=1}^{T}\sum_{c=1}^{C} \log P(y_t^{(c)}|s_t^{(c)})\right] \\
&= \sum_{c=1}^{C}\sum_{i=1}^{m_c} \mathtt{E}_{s_1^{(c)}}\left[\delta(s_1^{(c)}, i)\right] \cdot \log \pi_i^{(c)} + \\
&\quad \sum_{c=1}^{C}\sum_{c1=1}^{C}\sum_{i=1}^{m_{c1}}\sum_{j=1}^{m_{c2}}\sum_{t=1}^{T-1} \mathtt{E}_{s_t^{(c1)} s_{t+1}^{(c2)}}\left[\delta(s_t^{(c1)}, i) \cdot \delta(s_{t+1}^{(c2)}, j)\right] \cdot \log h_{i,j}^{(c1,c2)} + \\
&\quad \sum_{c=1}^{C}\sum_{i=1}^{m_c}\sum_{t=1}^{T} \mathtt{E}_{s_t^{(c)}}\left[\delta(s_t^{(c)}, i)\right] \cdot \log P(y_t^{(c)}|i) \\
&\triangleq \sum_{c=1}^{C}\sum_{i=1}^{m_c} \tilde{\pi}_i^{(c)} \cdot \log \pi_i^{(c)} + \\
&\quad \sum_{c=1}^{C}\sum_{c1=1}^{C}\sum_{i=1}^{m_{c1}}\sum_{j=1}^{m_{c2}} \tilde{\xi}^{(c1,c2)}(i,j) \cdot \log h_{i,j}^{(c1,c2)} + \sum_{c=1}^{C}\sum_{i=1}^{m_c} \tilde{\gamma}^{(c)}(i) \cdot \log P(y_t^{(c)}|i)
\end{aligned}
$$

According to the attributes of the expectation operator and the Kronecker delta operator, the sufficient statistics are given in the following way, and the parameters related to the state transitions are maximized by Equations 3 and 4:

$$
\begin{aligned}
\tilde{\pi}_i^{(c)} &= \gamma_i^{(c)} \\
\tilde{\xi}^{(c1,c2)}(i,j) &= \sum_{t=1}^{T=1} \xi_{t\to t+1}^{(c1,c2)}(i,j) \\
\tilde{\gamma}^{(c)}(i) &= \sum_{t=1}^{T} \gamma_t^{(c)}(i)
\end{aligned}
$$

The parameters are re-estimated in the same way as in the known latent state case.

Social Intelligence Design and Human Computing

Toyoaki Nishida

Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract. The central concern of Social Intelligence Design is the understanding and augmentation of social intelligence that might be attributed to both an individual and a group. Social Intelligence Design addresses understanding and augmentation of social intelligence resulting from bilateral interaction of intelligence attributed to an individual to coordinate her/his behavior with others in a society and that attributed to a collection of individuals to achieve goals as a whole and learn from experiences. Social intelligence can be addressed from multiple perspectives. In this chapter, I will focus on three aspects. First, I highlight interaction from the social discourse perspective in which social intelligence manifests in rapid interaction in a small group. Second, I look at the community media and social interaction in the large, where slow and massive interaction takes place in a large collection of people. Third, I survey work on social artifacts that embody social intelligence. Finally, I attempt to provide a structured view of the field.

Keywords: Social Intelligence Design, Conversational Informatics.

1 Introduction

The human is a social being. As an individual in society, a human should be able to manage relationships with other people and act wisely in a situation governed by an implicit or explicit set of social rules, in order to live a comfortable life. Such social skills range from an ability to communicate with and to learn from others to that of achieving goals by cooperating or negotiating with other people under a given social structure.

In the meanwhile, activities of humans constitute the human society. A collection of people may be considered to be intelligent if they can manage complexity and learn from experiences as a whole. The major concern of an organization is to maximize its problem solving power by coordinating the intellectual power of its members.

Social intelligence emerges from the synergy of individuals' social skill and the society's social structure. Both individuals' social skill and the society's social structure are indispensable to achieve social intelligence as a whole.

The advent of the network society has permitted people to interact with each other to affect social intelligence in a fashion unprecedented in history. On the one hand, it has brought about enormous benefits and conveniences. On the other hand, it has extended a dark side where a new technology is abused or disrupts human relations.

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 190–214, 2007.

The central concern of Social Intelligence Design is the understanding and augmentation of social intelligence. Researchers with scientific or sociological concerns might be interested in uncovering the mechanisms of social intelligence as a function of individuals' social skill and the society's social structure. Those with engineering minds might want to design and devise advanced artifacts that will contribute to amplifying social intelligence.

Social Intelligence Design becomes central to human computing whenever one sheds light on situations in which a human is interacting with other people. Social Intelligence Design addresses a wide spectrum of phenomena concerning humans and their socially intellectual activities, ranging from establishing mutual intentions with nonverbal communications to new social media on the Internet.

In this chapter, I describe the conceptual framework of Social Intelligence Design, and present an overview of work on Social Intelligence Design presented at the Social Intelligence Design workshops.

2 Approaches to Social Intelligence Design

Social intelligence results from the bilateral interaction of intelligence attributed to an individual to coordinate her/his behavior with others in a society and that attributed to a collection of individuals to achieve goals as a whole and learn from experiences.

Social intelligence is ubiquitous in our daily life. Consider the following scenario.

> Scenario 1: K and J were coming to Kyoto for sightseeing. Although they found some nice famous places by Web search, they decided to go to a less famous, but quieter place with atmosphere that was recommended by local people. They posted their experience on the weblog.

Social Intelligence manifests in a variety of ways. It is evident that K and J relied on the experiences of other people on choosing the place for sightseeing. Recommendation and reputation were used as a primary means of knowledge resource to make a decision, in addition to catalog information about points of sightseeing spots. They might have preferred "local people" as a profile of the information source concerning the local affairs. They not only utilized the social information but they also contributed their own experience to the society. We also tend to rely on social heuristics such as the flow of people or a cluster of restaurants (such as the more restaurants, the more competitive and the nicer, or the longer history, the more reputation and the nicer).

In addition, there has been an enormous amount of tacit information "beneath the surface", such as believed to have happened at the nonverbal levels (gestures, facial expressions, voice tone, dressing, posture, etc), that are not explicit in the text, if you imagined what happened, say, before, during and after the conversation between the two groups. For example, there might have been a scene such as Figure 1 in the discourse. K and J might have been walking on the stone steps and looking at some interesting scene after J referred K to something to the right by speech and eye gaze; K might have realized and have started looking to the right; K and J might, then, have shared a sense of good communication by having realized that they had achieved a

Fig. 1. A scene of a fall day in Kyoto (painted by Mayumi Bono)

joint attention and start a conversation; nobody else might have paid much attention to them or might have ignored them out of courtesy; all might have kept quiet so not to disturb the quiet atmosphere.

Now, consider another scenario.

> Scenario 2: T became a member of a new company. Soon after T came into the new office, he wanted to print his file. When he went to a printer room, he found a beautiful output was printed from a printer. He tried to set up parameters according to a manual, but he has not succeeded after many trials.

This scenario refers to an artifact (printer). Our modern life is surrounded by a huge number of artifacts of varying kinds. Although a naive thought about the introduction of artifacts is that they bring about great conveniences in our daily life, they may also result in a new stress. For example, we have to undertake various kinds of interaction with the artifact (e.g., setting parameters or issuing commands) in order to achieve a goal (e.g., having the content of a file printed on a paper), which may cause panic and loss of time, as illustrated in the above scenario.

In general, human-computer interaction induces social interactions. For example, the designer of the printer may embed the tips in addition to a normal usage manual. The tips may be implicit in the sense that it is implemented in a tacit way such as a shape or the light embedded in a button. The owner of a system may want to investigate customers' preferences by trying to infer usage patterns by computing statistics of the log data. In contrast, the user may communicate her/his message by actions such as choosing the product, answering customers' inquiries, or complaining. We can consider that artifacts are considered to be social communication media through which people exchange messages with each other or communicate with each other.

The social interaction features induced by an artifact may become more prominent as artifacts become more sensible and proactive to the user, autonomous to cope with novel situations, and capable of acting on behalf of the designer or owner. If this

enterprise is successful, the artifacts may well complement the incompleteness of human's ability and knowledge, and any flaw of the current user support for complex machinery, as suggested in Scenario 2, may be dissolved. It will lead to the notion of artifacts as embodied social media.

Social Intelligence Design is a research field aimed at the understanding and augmentation of social intelligence. In what follows, I will overview work in Social Intelligence Design mainly published in the previous Social Intelligence Design workshops[1] from three perspectives suggested above. First, I will focus on the interaction from the social discourse perspective in which social intelligence manifests in rapid interaction in a small group. Second, I look at the community media and social interaction in the large, where slow and massive interaction takes place in a large group of people. Third, I survey work on social artifacts that embody social intelligence. Finally, I will attempt to provide a structured view of the field.

3 Interaction in Social Discourse

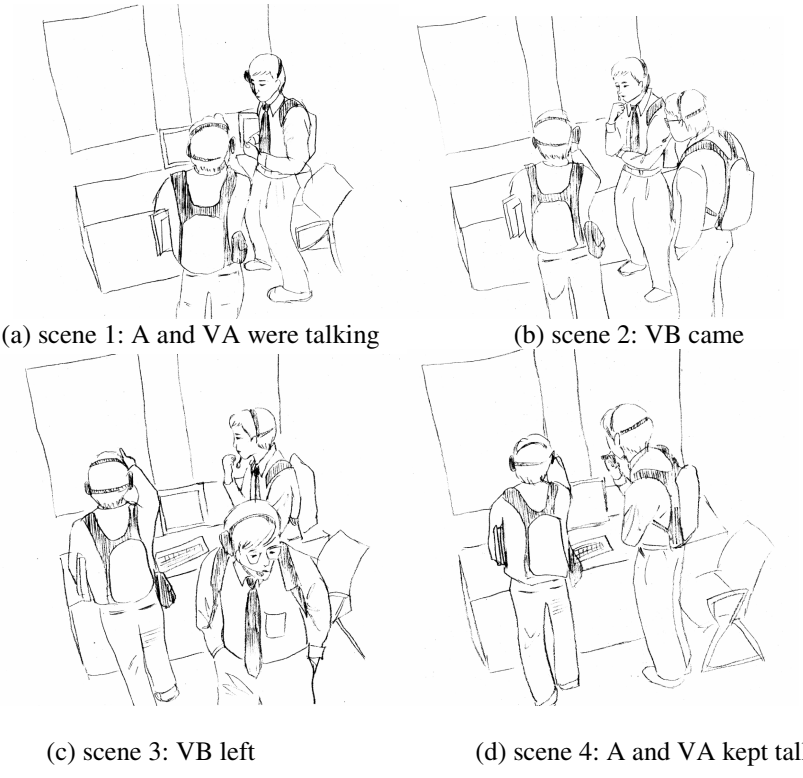
This perspective sheds light on structured interaction that takes place at the seconds-minutes order, at the conscious level. Collaboration protocol and story telling are key concepts at this level.

3.1 Social Interaction with Nonverbal Communication Means

People use nonverbal communication means in a proficient way to dynamically express their feelings, attitudes, intentions in addition to their thoughts. There is a tremendous amount of study on nonverbal communication behaviors including eye gaze, eyebrow, voice, facial expressions, head movement, hand gesture, body gesture, posture, distance, and so on. Kendon gave a structural description of gesture, most notably from the viewpoints of communication [Kendon 2004]. For example, he described in detail how greetings are made as a result of nonverbal interaction between social actors. McNeill studied relationships between gesture and thought [McNeill 05]. In addition to the development of measuring and coding schema for gestures, he presented a theory to characterize thought processes underlying gesture. He introduced a notion of growth point as "the smallest unit of the imagery-language dialectic", out of which a dynamic process of organization emerges. He characterized a growth point as "an empirically recoverable idea unit, inferred from speech-gesture synchrony and co-expressiveness." In addition, he introduced the notion of catchment to model multiple gestures with recurring form features to characterize the discourse segment to which the growth point belongs.

Some of these behaviors are intentionally produced by a communication partner for a communicative purpose, while others, such as a subtle correlation of eye gaze and mouth move, are not. In face to face communication, humans are considered to evaluate with each by sensing such unconscious, uncontrollable nonverbal behavior. There are several games such as poker where players are motivated to tell lies. Ueda and Ohmoto prototyped a real-time system that can discriminate lies by measuring gaze directions and facial feature points [Ohmoto 2006]. The system can measure

[1] http://www.ii.ist.i.kyoto-u.ac.jp/sid/



(a) scene 1: A and VA were talking (b) scene 2: VB came

(c) scene 3: VB left (d) scene 4: A and VA kept talking

Fig. 2. When the third person did not interrupt/join in an ongoing conversation (by courtesy of Dr. Mayumi Bono)

gaze directions and facial feature points, while allowing the user to move head position and orientation during the measurement and without requesting the user to place one or more markers on the face or preparing a face model in advance.

The discourse and task level information needs to be taken into account to see the role and function of each piece of communication behavior in the context of the entire discourse or task. Sacks introduced a concept of a turn-taking system in conversation, which suggests rules governing social interactions in conversation [Sacks 1974]. Multi-party face-to-face conversations such as those taking place at poster presentations are interesting as a subject of investigating mutual dependency between the task structure and the nonverbal communication behaviors. In the theory of participation status and framework, Goffman introduced such notions as side participants, bystanders, eavesdroppers in addition to the speaker and the addressee, in order to analyze the behavior of actors related to a conversation [Goffman 1981]. Bond used a framework of planning developed in Artificial Intelligence to model social behaviors as a set of joint steps with temporal and causal ordering constraints. An example of a joint behavior is grooming a pairwise affiliative behavior, which consists of four phases: orient, approach, prelude, and groom [Bond 2002].

Bono et al studied the dynamics of participation in face-to-face multi-party conversations [Bono 2004]. Figure 2 shows an example of a series of scenes at the poster session they investigated. In scene 1, an exhibitor (E) and a visitor (VA) were talking. During the talk, another visitor (VB) came in scene 2. Although VB came very close to E and VA in scene 3, VB left after around 27 seconds without exchanging conversation with A or VB. E and VA kept talking in scene 4.

It might be interesting to see the relationship between the utterances, distribution of eye gazes, and the arrangement of bodies. As for the distribution of eye gazes, E's eye direction was directed to VA, while VA was also gazing E in scene 1. Turn taking took place at an appropriate frequency. While one was making utterances as a speaker, he was looking at the hearer, making an explicit role assignment to the hearer. This interaction was maintained after VB arrived and VB was not chosen as a recipient of the utterance. As a result, VB had to stay as a side participant. This example scenario suggests that the role of conversational partner is determined proactively by the speaker.

Bono et al uncovered that the centrality of participation in a conversation is correlated with participants' interest in objects discussed as the topics of the conversation and that the more central a role one plays in a conversation, the more highly interested one is in the topic of conversation. In addition, they found that the frequency of interaction with exhibitors was a good indicator of visitors' interest in the corresponding posters, interest in the posters was proportionate with the time visitors spent being directly addressed by exhibitors rather than as side-participants, and addressee-hood of visitors manifested itself in the coordinated production of verbal backchannel responses.

Introduction of communication media may change the nature of communication. Mark and DeFlorio investigated how the use of wall-size, life-size HDTV was useful in overcoming the problems found in interaction with regular video-conferencing systems. The HDTV (High Definition TV) provides high-resolution images, conveyed to a far wider angle of the remote room with less distortion, than normal video-conferencing. Telephones were provided in each room, with telephone numbers, to support small group discussions (sidebar conversations) between the sites. They reported that people did not use exaggerated gestures or movements to convey expression through HDTV image in videoconferences. They also found that far fewer sidebar conversations across the geographically distant rooms were made during the videoconferences than during face-to-face conferences in the same room. They considered that the HDTV was mainly used for supporting the public conversations and as a result the team leader became a single primary channel of information for the group, as opposed to normal meetings, [Mark 2001]. In addition, they considered that the video image functioned as an awareness mechanism for activity in the other room, too.

3.2 Knowledge in Action

Gesture can be seen as external representations of abstract concepts which may be otherwise difficult to illustrate, as pointed out by Biswas and Fruchter [Biswas 2005]. They presented a framework for processing the captured video data to convert the tacit knowledge embedded in gestures into easily reusable and computable semantic

representations. They prototyped a system called I-Gesture to investigate how people use gestures to express ideas. I-Gesture allows users to define a vocabulary of gestures for a specific domain, build a digital library of the gesture vocabulary, and mark up entire video streams based on the predefined vocabulary for future search and retrieval of digital content from the archive.

In order to capture and mine social communicative events for further knowledge reuse, Yin and Fruchter developed a prototype called I-Dialogue that captures the knowledge generated during informal communicative events through dialogue in the form of an unstructured digital design knowledge corpus. In addition, I-Dialogue adds structure to the unstructured digital knowledge corpus, and processes the corpus using an innovative notion disambiguation algorithm in support of knowledge retrieval [Yin 2007].

Skilled cooperative actions embody knowledge in co-action and can be seen as a form of social intelligence for sustainable interaction. Gill and Borchers considered people, tools, artifacts, and technologies to be dynamic representations of knowledge in joint design activities and aimed to study these in co-action [Gill 2003]. They observed and compared the behaviors of students to collaborate to design shared dorm living spaces using a normal whiteboard and two SMARTboards (electronic whiteboards that allowed for web browsing, typing up session notes, drawing mockups, and viewing their work). Body Moves was coded as the Parallel Coordinated Move (PCM). They obtained useful insights into the use of surfaces in collaboration. They reported that the ability to engage at the surface resulted in different strategies depending on the nature of the surface. For example, on the SMARTboard surface, body moves such as take-turn were used. As a result, the body field of the person acting was disturbed when the other one was entering it, and a reconfiguration of the engagement space was required. At the whiteboard surface, body moves such as attempt contact and focus were used to increase contact. In addition, further interesting phenomena were observed where participatory structure dynamically changed according to the problem solving status. They called it interactional dance.

3.3 Meeting Capture

Synthetic approaches at this level can be applied to the meeting room applications in order to better support meetings, such as real-time browsing, retrieval and summarization of meetings. Intelligent support of meetings can be characterized in the context of ambient intelligence defined as ubiquitous computing plus social and intelligent interfaces [Nijholt 2005b]. Issues in designing a meeting environment involve interpretation of events and activities in the environment, provision of real-time support, multimedia retrieval and reporting, autonomous and semi-autonomous embodied agents, controlling the environment, and its inhabitants. The M4 (multi-modal meeting manager) project was aimed at designing a meeting manager that can translate the information captured from microphones and cameras into annotated meeting minutes that allow for high-level retrieval questions, and for summarization and browsing. In order to track discussions, group actions such as presentations,

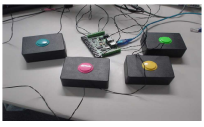


Fig. 3. Button-devices for human-assisted capture of conversation quanta [Saito 2005]

discussions, consensus and note-taking are modeled using Hidden Markov Models (HMMs). The actions of the individuals are recognized independently, and fused at a higher level. On the other hand, the augmented multi-party interaction (AMI) project aimed to develop new multimodal technologies to support human interaction in smart meeting rooms and remote meeting assistants. It aimed to enhance the value of multimodal meeting recordings and to make human interaction more effective in real time.

Saito et al reported a method of human-assisted conversation quanta acquisition using auxiliary button devices, as shown in Figure 3. In this conversation capture environment, each conversation participant is expected to press her/his button when s/he expects or finds a segment of conversation useful and hence intends to record the segment [Saito 2005]. The conversational situation itself is recorded during the conversation. Each participant can press the button to indicate the system her/his intention to mark. S/he can specify the in-point by quickly pressing the button n times to go back in the discourse $n \times d$ seconds, where d is a time unit for going back. Then, s/he keeps hold of the button until s/he thinks it is enough. The out-point will be set when s/he releases the button. As a result of an experimental investigation, five seconds was found to be an optimal value for d.

A meeting assistance environment called SKG Room as shown in Figure 4, was implemented. In this environment, a sphere surface of a spatial content aggregator called SKG is projected on a large screen. The users can have a conversation in front of the large screen. The conversation content can be captured by the human-assisted conversation contents capture environment. The captured conversation quanta are integrated on the SKG surface. In addition to the button devices, the touch screen was made available as a capture device, so that the user can touch the screen to signal her/his intentions.

In order to support the meeting in a more intelligent fashion, one might be interested in introducing pro-active meeting assistants that aim to assist the meeting process and thereby facilitate more effective and efficient meetings. Rienks et al discussed the requirement specifications of such meeting assistants and carried out a Wizard of Oz experiment [Rienks 2006]. A set of four different systems with varying intrusiveness levels was devised for the experiment. They were evaluated by two student committees of eight and seven members respectively. As a result, two versions that use voice samples together with a clock display when an item is due to be finished, something is off topic, a subject takes too long, or a discussion is unbalanced resulted in much more efficient meetings than those without pro-active meeting assistants, though less enjoyable.

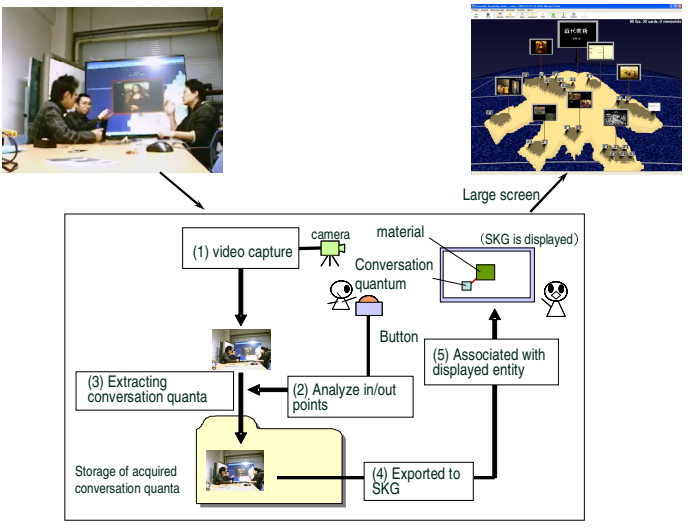


Fig. 4. SKG Room [Saito 2005]

3.4 Collaboration Support

Another key issue at this level is collaboration design. Fruchter proposes the analysis of collaboration support systems with three perspectives, physical spaces (e.g., "bricks"), electronic content ("bits"), and the way people communicate with each other ("interaction") [Fruchter 2001]. Fruchter introduced Brick & Bits & Interaction Hypothesis and Change Hypothesis and emphasize in designing their importance in collaboration technologies as follows:

Brick & Bits & Interaction Hypothesis: If we understand the relationship between bricks, bits, and interaction we will be able to

1. design spaces that better afford communicative events,
2. develop collaboration technologies based on natural idioms that best support the activities people perform, and
3. engage people in rich communicative experiences that enable them to immerse in their activity and forget about the technology that mediates the interaction.

Change Hypothesis: Any new information and collaboration technology will require change and rethinking of:

1. the design and location of spaces in which people work, learn, and play.
2. the content people create in terms of representation, media, interrelation among the different media, the content's evolution over time so that it provides context and sets it in a social communicative perspective.
3. the interactions among people in terms of the individual's behavior, interaction dynamics, new communication protocols, collaboration processes; relation between people and affordances of the space; and interactivity with the content.

The hypotheses were used in the Global Teamwork in Architecture, Engineering, Construction (A/E/C) offered at Stanford University.

4 Community Media and Social Interaction in the Large

This level is concerned with massive interaction in a society. It is slow and sometimes proceeds without much attention.

4.1 Understanding Community Media

There are several methodologies for analyzing macro interaction, involving social survey to collect information from the society, sociological methodology to interpret phenomena, social psychological methods, or statistic methods to derive a tendency. Miura and Shinohara investigated the communication congestion which is defined as the phenomenon of multiple topics simultaneously running when many participants are participating in the same chat session at the same time. [Miura 2004]. According to their study, high-density congestion can be regarded as leading to overloading of the participants' cognitive systems, making their remarks become less informative; moderate-density congestion with a relevant topic can be regarded as an activator of communication, particularly for experienced participants. Their research suggests that it might be possible to enhance the creativity of a group by adaptively changing the density of congestion.

A sociological approach may be applied to workplace design to facilitate communication and collaboration between mobile workers and their office base by ensuring compatibility between fixed and mobile, local and remote work areas. Rosenberg et al presented work on interaction space as the relationship between people, location, work task and process, based on the assumption that people jointly create an interaction space in which they work together towards achieving mutual understanding, and that the common ground is created in the interaction space that is shaped by spatial and organizational constraints as well as informational resources that determine the nature of the hybrid workplace. The investigation focused on the distance between people from three key perspectives [Rosenberg 2004].

In order to define a framework of standardized quantitative measurement of social intelligence, a notion of Social Intelligence Quantity (SIQ) was defined by combining the qualitative evaluation consisting of questionnaire and protocol analysis, and the quantitative evaluation consisting of the network log analysis, the factorial experiment and the standardized psychological scale [Yamashita 2002]. SIQ consists of SIQ-Personal and SIQ-Collective.

SIQ-Personal specifies the individual's personal attitudes to the society. SIQ-Personal is measured with the individual's information desire and intention to participate in the community. Matsumura investigated in detail the information desire and identified that the information acquisition desire consists of the interpersonal relation desire, the trend information acquisition desire, the information publication desire, the information monopoly desire, and the information acquisition desire [Matsumura 2004]. In order to investigate the structure of the individual's intention to participate in the community, Matsumura introduced seven factors (i.e., the intention

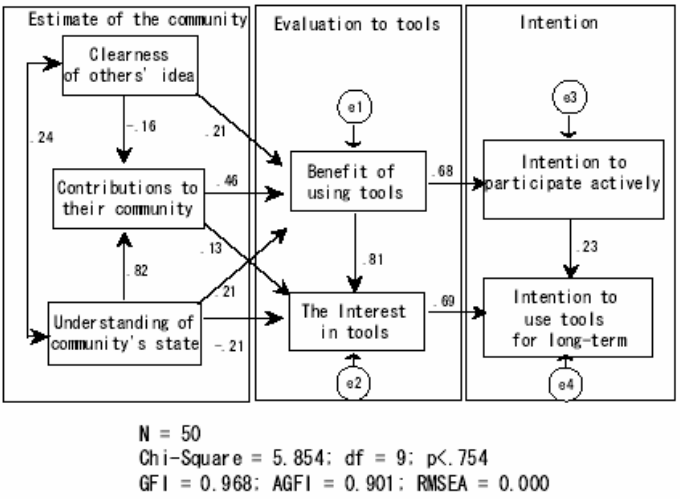


Fig. 5. A causal model concerning the intention of participating in a community and using a tool [Matsumura 2004]

of active participation, the intention of continuous participation, the benefits of using tools, the interest in the tool, the clarity of others' idea, the contribution to their community, and the understanding of the community's status), and proposed a causal model as shown in Figure 5. It indicates that the intention to participate in a community was influenced by the benefit received by using a communication tool. The benefit was affected by factors associated with understanding the state of a community. The individual's subjective contribution had a strong effect on the benefits of using a tool. The individual's subjective contribution depends on the understanding of the state of a community and the clarity of others' opinion. The model suggests that one should examine not only users' subjective evaluation of a communication tool but also their subjective evaluation of a community, in order to evaluate a communication tool for supporting a community.

SIQ-Collective represents the community's status of information and knowledge sharing or knowledge creation. Elaboration of SIQ-Collective is in progress. It is expected that SIQ-Collective may be defined by combining the amount of information in the community and the objective indices such as the diversity and the convergence of information.

4.2 Supporting Community

Digital Divide was addressed by Blake and Tucker. They discuss field trials that are underway in remote rural regions in South Africa in Tele-Health projects [Blake 2004]. There are a certain amount of synthetic techniques for producing awareness information by collecting information about what others are doing. FaintPop [Ohguro 2001] is a nonverbal communication device in which small icons of the user's colleagues are displayed. It allows the user to communicate her/his feeling towards her/his colleagues by using three types of touching. Fukuhara et al proposed a method

called temporal sentiment analysis for finding sentiment of people on social problems according to the timeline. It provides a topic graph on a sentiment category specified by a user, and a sentiment graph on a topic specified by him or her. From these graphs, one can find socially concerned topics from the viewpoint of sentiment, understanding how people thought about social events.

IBM's Babble is a large scale discussion support system based on the idea of social translucence [Erickson 2007]. Erickson characterizes social intelligence as the ways in which groups of people manage to produce coherent behavior directed towards individual or collective ends. He introduced a device called the social proxy which is a minimalist graphical representation that portrays socially salient aspects of an online interaction, such as aspects of presence or other social activities of people. Based on experiences with implementing a number of design experiments, Erickson made six claims about designing with social proxies: everyone sees the same thing; no user customization; portray actions, not interpretation; social proxies should allow deception; support micro/macro readings; ambiguity is useful: suggest rather than inform; and Use a third-person point of view.

4.3 Knowledge Circulation in a Community

Circulating knowledge in a community is critical to the community not only because it enables knowledge sharing and formation of the common culture but also it increases the value of knowledge by going through an evolutionary process. In order to facilitate the knowledge circulation process in a community, we need to reduce overheads for knowledge circulation. We have developed a suite of technologies for conversational knowledge circulation. It consists of human-assisted acquisition of conversational contents, visual accumulation of conversational contents, and content-driven embodied conversational agents.

Conversation quantization was introduced as a conceptual framework underlying the entire package (Figure 6). The central idea in conversation quantization is to use a conversation quantum that encapsulates interaction-oriented and content-oriented views of a conversation scene. The interaction-oriented view sheds light on how each conversation partner coordinates her/his behaviors to communicate with the other. The content-oriented view, on the other hand, focuses on how meaning emerges from interaction.

Consider a conversational situation illustrated on the left-hand side of Figure 6, where one person on the left is explaining to another person on the right how to detach the front wheel from a bicycle. The interaction-oriented view of this scene is to see how participants interact with each other. It is quite probable that while the person on the left is explaining the operation with verbal and nonverbal modalities, the person on the right may also react to the explanation, say by looking or nodding.

A content-oriented view of this scene might be a proposition "in order to detach the front wheel from the bicycle, one must turn the lever to the right." Alternatively, it might simply be another proposition such as "A is talking to B," or "A is helpful." Thus, a content-oriented view may give a high-level, conceptual description of a conversation scene, while the interaction-oriented view grounds a content-oriented view. The interaction and content oriented views constitute a conversation quantum, as suggested on the right hand side of Figure 7.

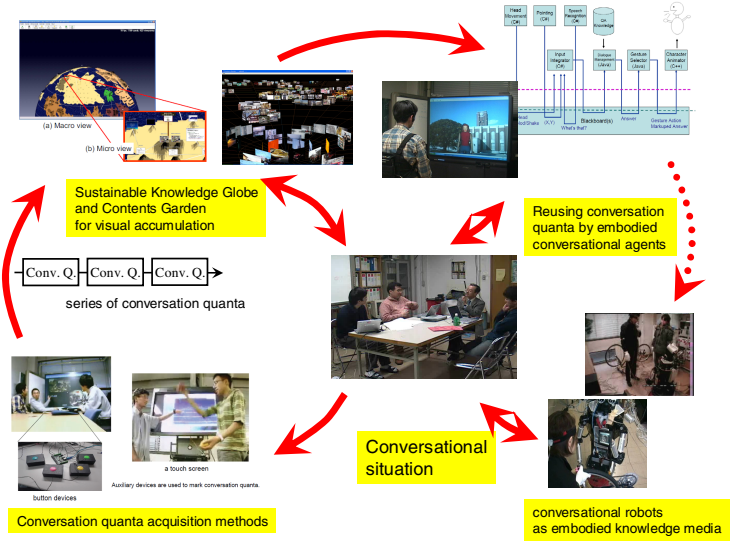


Fig. 6. Conversational Knowledge Circulation Enhancement Package

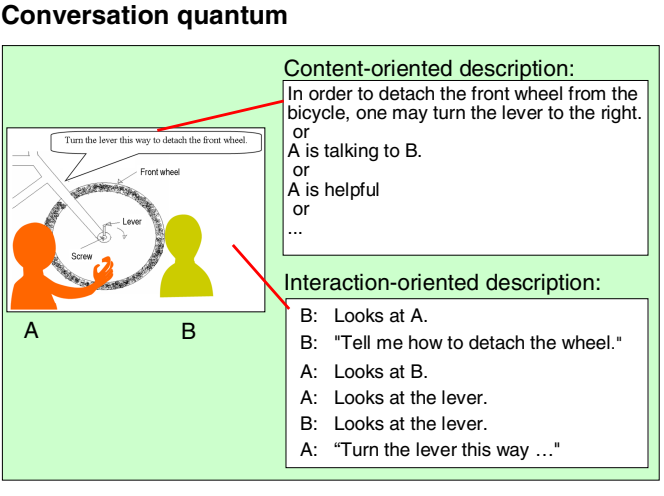


Fig. 7. A conversation quantum for a conversation scene

The scope of a conversation quantum is limited to a minimal conversation scene. This is because we would like to make a conversation quantum versatile so that it can be embedded in multiple conversation situations.

Capturing conversational contents from conversation scenes (materialization) involving identification of conversational scenes, recognition of communicative interactions and interpretation of communicative interactions is a challenging problem and hard to automate based on the current technology. A method of human-assisted

conversation quanta acquisition using auxiliary button devices was developed. In order for a user to benefit from a large collection of conversation contents, a long-term relationship between the user and the conversation contents need to be established so that the conversation contents may coevolve with user's biological memory. Towards this end, we believe that supporting the spatial and temporal evolution of visually accumulated conversation contents ("knowledge landscape") is effective. Content-driven embodied conversational agents are used as a user interface. In the future, communicative robots will be used as a real world interface.

5 Social Artifacts

Insights obtained from Social Intelligence Design may be implemented as a social artifact such as an ECA or a conversational robot that can undertake social interaction with other social actors. (or -- that can interact socially)

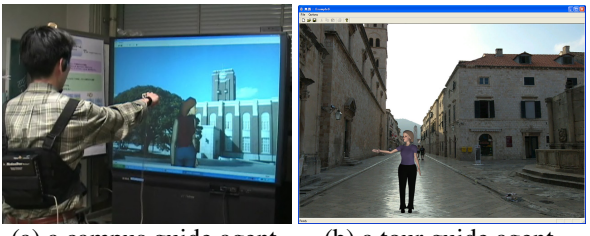
5.1 Embodied Conversational Agents

Embodied conversational agents (ECAs) are human-like characters that can interact with the user in a conversational fashion. As such, they are expected to induce a stronger form of the media equation [Reeves 1996] and to allow for natural social interaction with the user. In order to design an ECA that can mediate inter-human communication, we need to consider many aspects including ease of content management, a virtual environment that provide the context and referents for conversation, and implementation of higher order social concepts such as politeness, friendliness, humor, personality or trust.

Nijholt proposed an information-rich virtual environment mimicking a real theater building where the user interacts with autonomous agents [Nijholt 2001]. The mental model of each autonomous agent is modeled in terms of beliefs, desires, plans, and emotions.

GECA (Generic ECA) is a platform (Figure 8) that can execute an immersive ECA system on multiple servers connected with each other by a computer network [Huang 2006]. GECA features:

- A general purpose platform (GECA platform) and a set of managing programs for mediating and transporting data stream and command messages between stand-alone ECA software modules that compose an ECA system interacting with human users in multiple modalities.
- A specification of a high-level protocol (GECAML) based on XML messages that are used in the communication between a standardized set of ECA components such as sensor inputs from the human users, inference engine, emotion model, personality model, dialogue manager, face and body animation, etc.
- An application programming interface (API) available on main-stream operating systems and programming languages for easily adapting ECA software modules to hook to the platform.



(a) a campus guide agent (b) a tour guide agent

(c) a quiz agent (d) a pedagogical agent

Fig. 8. ECAs built on top of GECA [Huang 2006]

The blackboard model is employed as the backbone to realize the weak inter-connectivity of the components, enabling the online switching of components and offline upgrading and maintaining of components. It allows logically isolated multiple blackboards to distribute information traffic concentrated on a logically single blackboard.

GECA has been implemented and applied for various applications involving a campus guide agent, a tour guide agent, a quiz agent, and a pedagogical agent for teaching cross cultural communication (Figure 8).

5.2 Communicative Robots

Understanding the functional and emotional aspects of nonverbal interaction is a key issue. Not all nonverbal behaviors are consciously controlled by the actors. By measuring the behavior of actors, we can find interesting aspects of social intelligence. Xu et al. built a human-human WOZ (Wizard-of-Oz) experiment setting that allows for observing and understanding how the mutual adaptation procedure occurs between human beings in nonverbal communication [Xu 2007]. The motivation of the WOZ experiment was to observe the nature of mutual adaptation of human-human interaction, which was intended to be implemented on robots. Figure 9 shows the experiment setting. One human subject (the actor) was asked to move around in a simulated mine field to clean target objects according to the instruction of another human subject (the instructor). The instructor was given the entire task map with the positions of all the obstacles on the map, including bombs, signals, targets, as well as the goal and orbit. The instructor would also need to perceive the current, exact position of the actor. The reward, in the form of sound effects (clearing target objects, exploding bombs/signals, reaching the goal, etc.) and scores were necessary for both the instructor and the actor. The mask and sunglasses were used to prevent

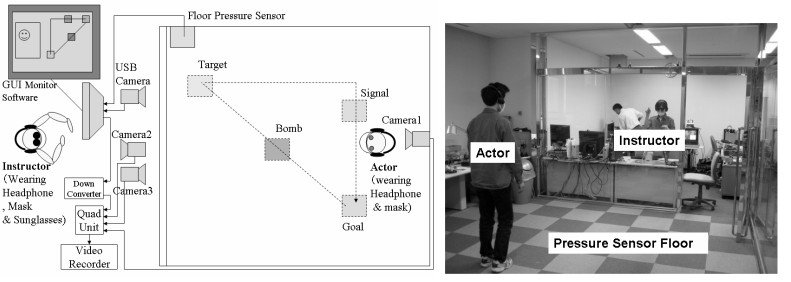


Fig. 9. A WOZ experiment setting for observing mutual adaptation [Xu 2007]

interference from other communication channels. The decision was made to prevent the subjects from communicating with eye gaze or facial expressions that were expected to be difficult to detect automatically when the observed behaviors would be implemented on robots. The instructor observes the movement of the actor by viewing the floor sensor information displayed on the computer screen. In order to enable the instructor to observe the actor's movements more clearly, a USB web camera was installed. As a result, three interesting findings were observed frequently. The first was called alignment-based action in which two participants align their actions while interacting with each other. The second was symbol-emergent learning in which the instructor used gestures such as symbol-like instructions when he interacted with the actor. The third was environmental learning in which communication between the participants became smoother in the subsequent trials and both participants adapted and improved their efficiency of movement.

5.3 Establishing Mutual Intention

Social intelligence at this level might be implemented as cohabiting with people. In order for a robot to cohabit with people, the robot should be able to be sensitive to subtle nonverbal behaviors of people. For example, in order for a robot to speak to a person who is making conversation with another person, it should follow a normal protocol people share in the society, that is, coming in the sight of the person, glance at her/his eye, wait for a call, and start talking as soon as it is called, just as is observed in inter-human communication [Kendon 1973]. The behavior should be modulated depending on how urgent the message is (Figure 10).

The key observation here is that the mutual intention formation between the client and the robots is established as a result of quick interactions using quick nonverbal interactions. The coordination of their behaviors is made by choosing options at each stage. If the gentleman does not want to be interrupted for a while, or the robot is called by some other urgent requests, they will act differently at respective stages, resulting in a different outcome.

We consider it is feasible to implement artifacts that can communicate with people with nonverbal communication means. The ability of forming and sustaining intention shared by participants (joint intention) is considered to be a primary goal in

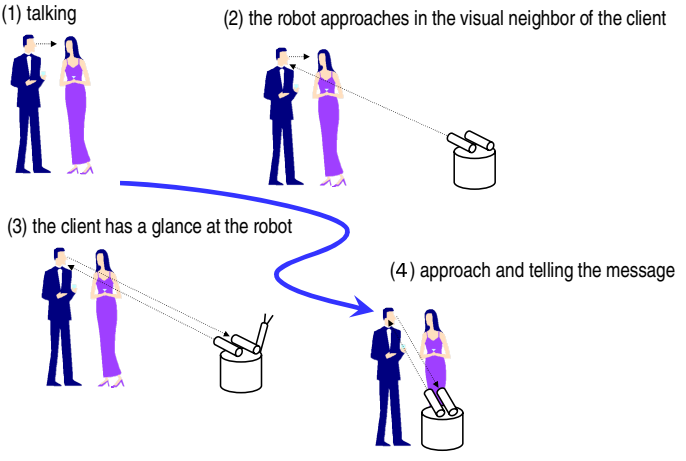
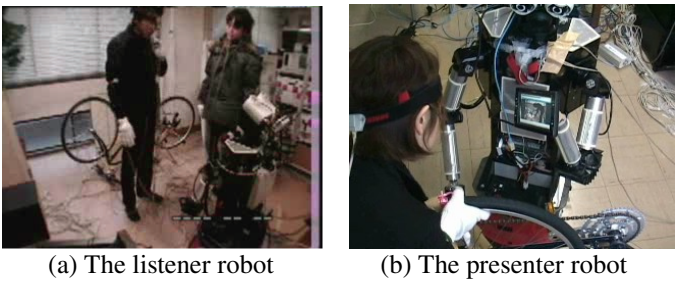


Fig. 10. How conversational robots should behave to speak to a person in conversation

nonverbal communication. The communication schema we have in mind allows two or more participants to repeat observations and reactions at varying speeds to form and maintain joint intentions to coordinate behavior, which may be called a "coordination search loop" [Nishida 2006].

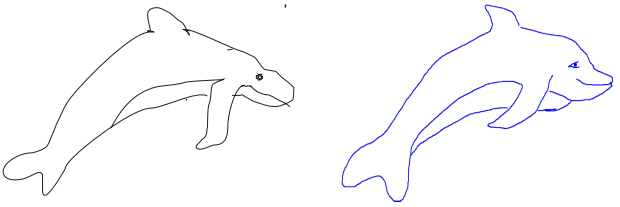
A study on embodiment communication with a mobile chair robot (a locomotive chair that can dynamically produce a means of allowing a person to get a place to sit down) suggests that although the users were all able to sit down on the chair as a result of coordinating behaviors, some users pointed out that the autonomous mobile chair should have communicated its intentions more explicitly [Terada 2001].

Listener and presenter robots were built by taking into account the lessons learnt from the mobile chair robot (Figure 11) [Nishida 2006, Ohya 2006]. The motivation underlying this research is to build a robot that can mediate knowledge among people. The listener robot interacts with an instructor to acquire knowledge by videotaping important scenes of her/his activities (e.g., assembling/disassembling a machine). The presenter robot, equipped with a small display, will then interact with a novice to show



(a) The listener robot (b) The presenter robot

Fig. 11. Robots as embodied knowledge media



(a) Initial drawing (b) Final drawing

Fig. 12. Is this what you intended to draw? [Mohammad 2007a]

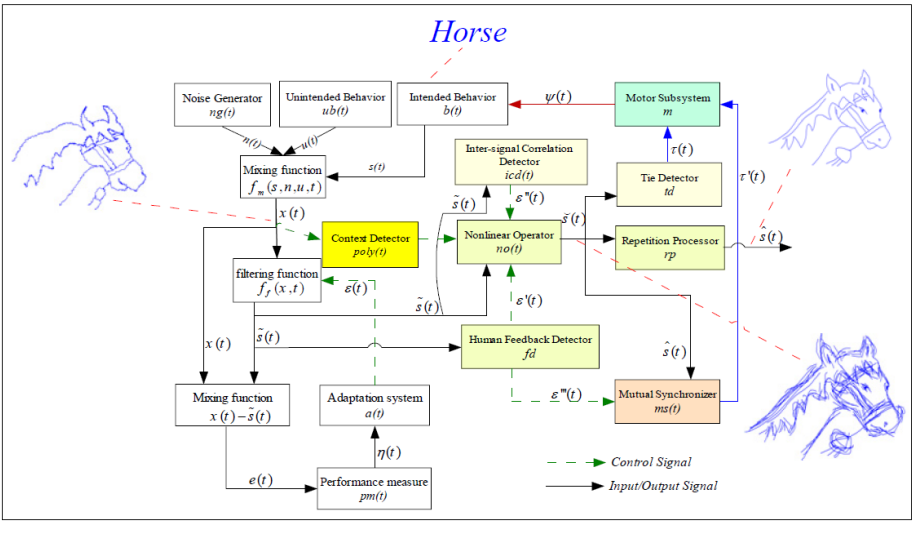


Fig. 13. The architecture of NaturalDraw [Mohammad 2007b]

the appropriate video clip in appropriate situations where this knowledge is considered to be needed during her/his work (e.g., trying to assemble/disassemble a machine).

From time to time, intention may not be clear, but it may become gradually concrete through interaction. A good example is drawing. Figure 12a and b were drawn by the same person using drawing software called NaturalDraw [Mohammad 2007a]. If the person was asked if the drawing in Figure 12a was what was wanted to draw, the answer would be no, for it was revised as in Figure 12b.

The idea of "coordination search loop" between the human user and the drawing software also helps here to derive the final drawing as a mutual intention between the user and the software [Mohammad 2007a]. Mohammad and Nishida used signal processing techniques based on the idea that intention can be best modeled not as a fixed unknown value, but as a dynamically evolving function. Interaction between two agents couples their intention functions creating a single system that co-evolves as the interaction goes toward a mutual intention state. They distinguish the intended signal from noise and unintended signal, and how the intended signal can be amplified through the repeated cycles of noise reduction and intention detection

including automatic repetition detection that enables the system to interactively detect the existence of repetition in the drawing, as shown in Figure 13. Roughly, it is assumed that the input to the system is a mixture of intended behavior, unintended behavior, and noise. The noise is filtered out by repeatedly applying filtering functions. Low pass filters are not used, for they may remove intended high frequency signals. In the meanwhile, an inter-signal correlation detector and a human feedback detector are used to extract intended behaviors such as repetition and the result is shown to the user, who is expected to strengthen or weaken the system's interpretation by the feedback.

6 A Structured View of Social Intelligence Design

In this section, I attempt to encompass the entire business of Social Intelligence Design in a more structured fashion. First, I give a historical overview of the field. Then, I introduce several viewpoints to classify approaches to Social Intelligence Design.

6.1 Historical Overview of Social Intelligence Design

The ideas of social intelligence design developed along with a series of international workshops starting in 2001.[2]

The first workshop, SID2001, was intended to launch a new interdisciplinary study focusing on social intelligence, aimed at integrating understanding and development of new information and communication technology that may even result in the emergence of a new language and lifestyle [Nishida 2001]. The dual views of social intelligence, an individual's ability to be able to "manage relationship with other agents and act wisely in a situation governed by an implicit or explicit set of shared rules, based on an ability of monitoring and understanding of other agents' mental state" and an ability of a group of people to manage complexity and learn from experiences as a function of the design of social structure, were identified and equally emphasized.

The scope of Social Intelligence Design identified at SID2001 is summarized in Table 1. The primary concern of Social Intelligence Design was identified as design and implementation of novel communication means for mediating interaction among people and agents. The scope ranges from preliminary and preparatory interactions among people such as knowing who's who, to more intimate interaction such as collaboration. Supporting a group formation, collaboration, negotiation, public discussion or social learning is considered to be an important application of Social Intelligence Design. Theoretical aspects involve designing, deploying, and evaluating social intelligence support tools.

The second workshop, SID2003, focused on the theories to enhance the understanding and conceptualization of human cognition and interpersonal interaction to emphasize the practical aspects of Social Intelligence Design [Rosenberg 2005]. The major outcomes of the second workshop include: experimental approach to the

[2] http://www.ii.ist.i.kyoto-u.ac.jp/sid/

design of interactive workspaces, ethnographic studies of the introduction of technologies, the building of stories, methodologies for studying construction projects, the studies of online communities where mediated communication is a key vehicle for creating and maintaining social contact, building an interactive system that aids human decision and action, an analysis of the dynamism of the virtual community, and analysis of the effect of anonymous communication in online community.

The third workshop, SID2004, distinguished the following four themes [Nijholt 2006a]:

1. Natural Interactions - covering theory, modelling and analytical frameworks that have been developed with Social Intelligence Design in mind, including situated computation, embodied conversational agents, sociable artefacts, socially intelligent robots.
2. Communities - covering community media, communication patterns in online communities, knowledge-creating, network and anonymous communities.
3. Collaboration Technologies and tools - covering innovations to support interactions within communities, covering a range from knowledge sharing systems, multi-agent systems and interactive systems.
4. Application Domains - including design, workspaces, education, e-commerce, entertainment, digital democracy, digital cities, policy and business.

The major outcomes of the workshop involve the analysis of the effect of employing embodied conversational agents or communicative robots in human-artifact communication, analysis of presence in distributed workspaces, analysis of the effect of long-term use of distributed workspaces, analysis and modeling of meetings, analysis of network search activities, and an effort to compensate for the digital divide.

The fourth workshop, SID2005, added

- Multidisciplinary perspectives – exploring Social Intelligence Design at the intersection of different disciplines, such as, people-place-process, place-technology-interaction, that brings technology, work spaces, social behaviors and process aspects together.

to the fourth workshop [Fruchter 2007]. The major results involve a theory of improvisational social acts and communication, extending social intelligence design issues in extreme emergency and regular meeting room cases, analysis of the weblog community, supporting knowledge capture, sharing, transfer and reuse in collaborative teams.

The fifth workshop, SID2006, emphasizes the following three subjects:

1. Development, operation, and evaluation of support systems or tools for SID
2. Observation and modeling of psychological and behavioral processes of e-community
3. Social intelligence design by pilot program and computer-aided simulation.[3]

[3] http://www.team1mile.com/asarin/sid2006/

Table 1. The Scope of Social Intelligence Design identified at SID2001 [Nishida 2001]

1. Theoretical aspects of social intelligence design
   (a) Understanding group dynamics of knowledge creation
   (b) Understanding consensus formation process
   (c) Theory of common ground in language use
   (d) Attachment-based learning for social learning
2. Methods of establishing the social context
   (a) Awareness of connectedness
   (b) Circulating personal views
   (c) Sharing stories
3. Embodied conversational agents and social intelligence
   (a) Knowledge exchange by virtualized egos
   (b) Conversational agents for mediating discussions
   (c) A virtual world habited by autonomous conversational agents
   (d) Social learning with a conversational interface
   (e) Conversations as a principle of designing complex systems
   (f) Artifacts capable of making embodied communication
4. Collaboration design
   (a) Integrating the physical space, electronic content, and interaction
   (b) Using multi agent system to help people in a complex situation
   (c) Evaluating communication infrastructure in terms of collaboration support
5. Public discourse
   (a) Visualization
   (b) Social awareness support
   (c) Integrating Surveys, Delphis and Mediation for democratic participation
   (d) Evaluation of social intelligence
   (e) Network analysis
   (f) Hybrid method

The sixth workshop, SID2007, focuses on

1. Development, operation, and evaluation of support systems or tools for SID, which includes support systems and tools both for mediated remote interaction and support for face-to-face interaction;
2. Observation and modeling of psychological and behavioral processes with the aim of obtaining computational models of behavior and interaction;
3. Social intelligence implemented in interfaces, embodied agents, storytelling environments, (serious) gaming and simulation.[4]

6.2 Viewpoints to Classify Approaches to Social Intelligence Design

We might be able to classify approaches to Social Intelligence Design depending on which aspects of social intelligence are addressed.

(1) Micro in the small versus macro in the large

Social intelligence may manifest in a quick interaction among a small number of participants, as discussed concerning the first scenario. In contrast, it may be

[4] http://hmi.ewi.utwente.nl/sid07

observed in a slow interaction among a large number of participants. Typical examples are exchange of reputation and sharing of footprints on the Internet.

(2) Inter-human communication versus human-computer interaction

Social intelligence may be discussed in the context of inter-human communication. Computer-mediated communication falls into this category, where inter-human communication is mediated by ICT. In contrast, it may be discussed in the context of human-computer interaction, where a service reflects the intentions of various people including the owner, service provider, computer vender, and so on.

(3) Physical versus virtual

Social interaction takes place at the physical face-to-face interaction settings or in a virtual environment on the net. The phenomena may be more subtle and vague in the former case, for it is not clear exactly what features affect social intelligence.

(4) Speech acts

One might be interested in looking at the phenomena by looking at the speech acts involved in social interaction. Among others, cooperation, competition, and negotiation might be interesting. Higher level concepts include politeness or friendliness.

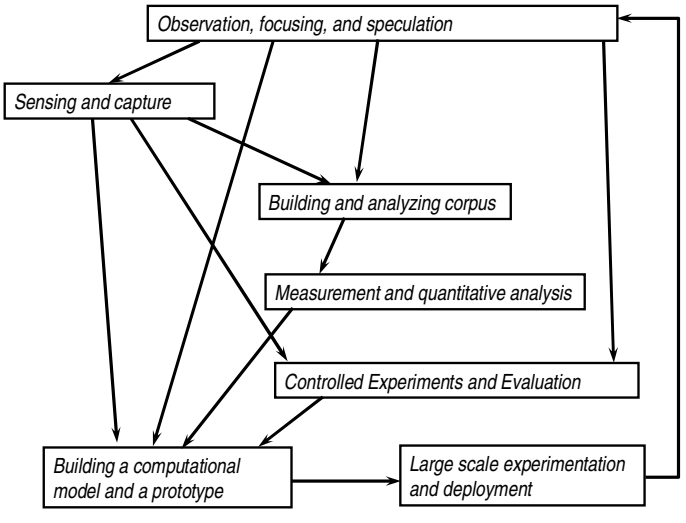


Fig. 14. Methodological Aspects of Social Intelligence Design

On the other hand, approaches to social intelligence design range from analysis motivated by scientific interests to synthesis motivated by engineering goals. Some might place much emphasis on understanding based on observation and investigation, trying to arguably identify and uncover critical aspects of social intelligence. Some others might be interested in modeling some aspects of social intelligence by following psychological disciplines. Yet others might be interested in building a

system that may augment collective intelligence of a group or a community. It should be noted, however, that an integrated approach is very important in Social Intelligence Design. Analysis and synthesis should be closely integrated. Analysis is needed to predict and evaluate applications. Synthetic approach is needed to focus analysis. Individual researches in Social Intelligence Design might be located and interrelated in a big picture as shown in Figure 14, depending on which aspects are emphasized as primary focus.

7 Concluding Remarks

In this chapter, I described Social Intelligence Design as an endeavor of understanding and augmentation of social intelligence that might be attributed to both an individual and a community. Social Intelligence Design involves understanding and augmentation of social intelligence, which is defined as an association of intelligence attributed to an individual to coordinate her/his behavior with others in a society and that attributed to a collection of individuals to achieve goals as a whole and learn from experiences. In this chapter, I first surveyed work on interaction in social discourse perspective in which social intelligence manifests in rapid interaction in a small group. Then, I looked at the community media and social interaction in the large, where slow and massive interaction takes place in a large group of people. Then, I surveyed work on social artifacts that embody social intelligence. Finally, I attempted to provide a structured view of the field.

References

[Azechi 2001] Azechi, S. Community based Social Intelligence, SID-2001, 2001.

[Biswas 2007] Biswas, P. and Fruchter, R. Using Gestures to Convey Internal Mental Models and Index Mutimedia Contents, AI & Society, to appear.

[Blake 2007] Blake E. and Tucker, W. User Interfaces for Communication across the digital divide, AI & Society, to appear.

[Bond 2002] Bond, A. H. Modeling Social Relationship -- An agent architecture for voluntary mutual control, in: Dautenhahn, K. Bond, A. H., Canamero, L. and Edmonds, B. (eds) Socially Intelligent Agents -- Creating Relationships with Computers and Robots, Kluwer Academic Publishers, pp. 29-36, 2002.

[Bono 2004] Bono, M., Suzuki, N., and Katagiri, Y. An Analysis of Participation Structures in Multi-Party Conversations: Do interaction behaviors give clues to know your interest?, Journal of Japanese Cognitive Science Society, No.11, Vol. 3, pp. 214-227, 2004.

[Erickson 2007] Erickson, T. 'Social' Systems: Designing Digital Systems that Support Social Intelligence. To appear in AI and Society.

[Fruchter 2001] Fruchter, R. Bricks & Bits & Interaction, Yukio Ohsawa, Shusaku Tsumoto, and Takashi Washio (eds): Exploring New Frontiers on Artificial Intelligence - Selected Papers from the First International Workshops of Japanese Society of Artificial Intelligence -, Lecture Notes on Artificial Intelligence LNAI2253, Springer Verlag, December 2001.

[Fruchter 2005] Fruchter, R., Nishida, T., Rosenberg, D. Understanding mediated communication: the social intelligence design (SID) approach, AI & Society, Volume 19, Number 1, pp. 1-7, 2005.

[Fruchter 2007] Fruchter, R., Nishida, T., Rosenberg, D. Mediated Communication in Action: a Social Intelligence Design Approach, AI & Society, to appear.

[Gill 2003] Gill, S. P. and Borchers, J. Knowledge in Co-Action: Social Intelligence in Using Surfaces for Collaborative Design Tasks, presented at SID-2003.

[Goffman 1981] Forms of talk. Philadelphia: University of Pennsylvania Press, 1981

[Huang 2006] Huang, H., Masuda, T., Cerekovic, A., Tarasenko, K., Pandzic, I., Nakano, Y., and Nishida, T., Toward a Universal Platform for Integrating Embodied Conversational Agent Components, in: B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part II, LNAI 4252, pp. 220 – 226, 2006.

[Kendon 1973] Kendon, A. and Ferber, A. A Description of Some Human Greetings, In: Michael, R. P. and Crook, J. H. (eds.) Comparative Ecology and Behaviour of Primates, Academic Press, 1973.

[Kendon 2004] Kendon, A. Gesture: Visible Action as Utterance. Cambridge: Cambridge University Press, 2004.

[Mark 2001] Mark, G. and DeFlorio, P. HDTV: a challenge to traditional video conferences? Publish-only paper, SID-2001, 2001

[Matsumura 2004] Matsumura, K. The Measures for the Evaluation of Communication Tools: the Causality between the Intention and Users' Subjective Estimation of Community, in Proceedings of the 3rd Workshop on Social Intelligence Design (SID 2004), pp. 85-90, 2004.

[McNeill 05] McNeill, D. Gesture and Thought, Chicago: University of Chicago Press, 2005

[Miura 2005] Miura, A. and Shinohara, K. Social intelligence design in online chat communication: a psychological study on the effects of "congestion", AI & Soc (2005) 19: 93–109.

[Mohammad 2007a] Mohammad, Y. F. O. and Nishida, T. NaturalDraw: Interactive Perception Based Drawing for Everyone, in Proc. 2007 International Conference on Intelligent User Interfaces (IUI 2007), pp. 251-260, 2007

[Mohammad 2007b] Mohammad, Y. F. O. and Nishida, T. Intention through Interaction: Toward Mutual Intention in Real World Interactions, IEA/AIE 2007, Kyoto, Japan (to be presented)

[Nijholt 2001] Nijholt, A. From Virtual Environment to Virtual Community, in Springer-Verlag LNAI 2253 New Frontiers in Artificial Intelligence (Joint JSAI 2001 Workshop Post-Proceedings), 2001.

[Nijholt 2006a] Nijholt, A. and Nishida, T. Social intelligence design for mediated communication, AI & Soc (2006) 20: 119–124, 2006.

[Nijholt 2006b] Nijholt, A., Op Den Akker, R., and Heylen, D. Meetings and Meeting Modeling in Smart Environment, Soc (2006) 20, 2006.

[Nishida 2001] Nishida, T. Social Intelligence Design -- An Overview, in: Takao Terano, Toyoaki Nishida, Akira Namatame, Yukio Ohsawa, Shusaku Tsumoto, and Takashi Washio (eds): Exploring New Frontiers on Artificial Intelligence - Selected Papers from the First International Workshops of Japanese Society of Artificial Intelligence -, Lecture Notes on Artificial Intelligence LNAI2253, Springer Verlag, December 2001.

[Nishida 2006] Nishida, T., Terada, T., Tajima, T., Hatakeyama, M., Ogasawara, Y., Sumi, Y., Xu Y., Mohammad, Y.F.O., Tarasenko, K., Ohya, T., and Hiramatsu, T., Towards Robots as an Embodied Knowledge Medium, Invited Paper, Special Section on Human Communication II, IEICE Trans. Information and Systems, Vol. E89-D, No. 6, pp. 1768-1780, June, 2006.

[Ohguro 2001] Ohguro, T., Kuwabara, K., Owada, T., and Shirai, Y. FaintPop: In Touch with the Social Relationships, in Springer-Verlag LNAI 2253 New Frontiers in Artificial Intelligence (Joint JSAI 2001 Workshop Post-Proceedings), pp. 11-18, 2001.

[Ohmoto 2006] Ohmoto, Y., Ueda, K., and Ohno, T. The real-time system of measuring gaze direction and facial features, and applying the system to discriminating lies by diverse nonverbal information, presented at SID-2006.

[Ohya 2006] Ohya, T., Hiramatsu, T., Yu, Y., Sumi, Y., and Nishida, T. Towards Robot as an Embodied Knowledge Medium―Having a robot talk to humans using nonverbal communication means, presented at SID-2006.

[Reeves 1996] Reeves, B and Nass, C. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, CSLI/Cambridge University Pres, 1996.

[Rienks 2006] Rienks, R., Nijholt, A. and Barthelmess, P. Pro-active Meeting Assistants: Attention Please!, presented at SID-2006.

[Rosenberg 2005] Rosenberg, D., Foley, S., Lievonen, M., Kammas, S., and Crisp, M. J. Interaction spaces in computer-mediated communication, AI & Soc (2005) 19: 22–33.

[Sacks 1974] Sacks, H., Schegloff, E., & Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. Language, 50 (4), 695-737, 1974.

[Saito 2005] Saito, K., Kubota, H., Sumi, Y., and Nishida, T., Analysis of Conversation Quanta for Conversational Knowledge Circulation, in: R. Khosla et al. (Eds.): *KES 2005*, LNAI 3683, pp. 296-302, 2005.

[Xu 2007] Xu, Y., Ueda, K., Komatsu, T., Okadome, T., Hattori, T., Sumi, Y., and Nishida, T. WOZ experiments for understanding mutual adaptation, AI & Society, to appear.

[Yamashita 2002] Yamashita, K, and Nishida, T. SIQ (Social Intelligence Quantity): Evaluation Package for Network Communication Tools, APCHI 2002 -- 5th Asia Pacific Conference on Computer Human Interaction - Beijing, China, 1-4 November 2002.

[Yin 2007] Yin, Z. and Fruchter, R. I-Dialogue: Information Extraction from Informal Discourse, AI & Society (to appear).

Feedback Loops in Communication and Human Computing

Rieks op den Akker and Dirk Heylen

Human Media Interaction Group, University of Twente, PO Box 217,
7500 AE Enschede, The Netherlands
infrieks@ewi.utwente.nl

Abstract. Building systems that are able to analyse communicative behaviours or take part in conversations requires a sound methodology in which the complex organisation of conversations is understood and tested on real-life samples. The data-driven approaches to human computing not only have a value for the engineering of systems, but can also provide feedback to the study of conversations between humans and between human and machines.

1 Introduction

An important aim for research that puts itself under headings such as Affective Computing, Ambient Intelligence, or Human-(Centered)-Computing, is to build systems that are able to interact with humans based on capabilities that are similar to those humans use to interact with each other. Being able to interpret human behaviour and determine the rational and affective concerns, motives and goals that lie behind it is a central capability of humans that are naturally inclined to take an intentional stance and have developed complex signalling systems to communicate their beliefs, intentions and to show or hide their attitudes and emotions.

With these central concerns, the research in such fields as Affective Computing can be seen to fit into the tradition of Artificial Intelligence and has in several respects complementary goals of research as Natural Language Processing. Whereas, NLP traditionally restricts its scope to language and the interpretation of utterances in semantic and pragmatic terms, Affective Computing focusses mainly but not exclusively on inferring information on the affective and the mental state of a person from physiological signals or forms of nonverbal communication.

For the creation of ambient intelligent systems a combination of approaches is needed that goes beyond the individual disciplines. Extending the work in NLP to include other modes of communication and other functional variables such as affect and attitude will also require rethinking methodology, theories and models. Similarly, the work in affective computing as it is currently practiced will need to re-orient itself if it is to be succesfully incorporated in a broader initiative that considers the full complexity of human communication. It is important

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 215–233, 2007.

for the success of visions such as ambient intelligence or affective computing to overcome the many restrictions that are self-imposed by disciplines or simplifications that are assumed due to the divide and conquer strategies that are common in scientific and engineering practice. One of the important methodological shortcomings to date is the reliance on non-naturalistic data that is studied out of context in many disciplines. In particular, affective computing suffers from this restriction as the summary of the state of the art and research programme as presented in [Pantic et al., 2007] shows. Moreover the theoretical background of the programme is based primarily on psychological studies that themselves study (affective) behaviours outside natural contexts of occurrence. In this paper, we sketch a complementary view that studies human interactions as they occur naturally as the basis for computational modelling and ambient applications.

Ultimately an ideal system would know how to deal with 1) the full gamut of human communicative behaviours, 2) the full range of meanings produced, and 3) the full complexity of human communication. Currently, the focus of the various research disciplines on one or more modalities, fails to address the intricate ways in which communicative behaviours are composed into complexes that are highly context dependent. The notion of communicative behaviour is often restricted to the typical 'expressive behaviours' such as language or non-verbal communication. However, intention abduction also takes place when someone observes another person simply carrying out an action. The focus on either semantic/pragmatic issues or on affective parameters ignores the various relations between them and aspects of the mental state of persons that are expressed through their behaviours such as propositional and interpersonal attitudes. The complexity of human communication shows, amongst others, in the different ways in which behaviours can express meanings that may shift depending on the context. For instance, facial expressions that first appear as symptoms of an emotional experience may shift to iconic expressions in a conversational context.

In several projects we have worked on the boundaries of disciplines or have attempted to move beyond them. In our work on affective dialogue systems, for instance, we explored the option to interpret facial expressions and other behaviours in the context of the ongoing dialogue using the appraisal checks as labels mediating between the contextualised action of the student on the one hand and the emotional appraisal on the other ([Heylen et al., 2005]). In the same context we looked at the generation of the appropriate responses of the dialogue act system taking into account the estimated affective state of the student that determined the general teaching strategy, the choice of dialogue act and the wording of the utterance ([Heylen et al., 2004]). In the AMI and AMIDA projects (Augmented Multiparty Interaction, http://www.amiproject.org)), we take human face-to-face conversations as the domain of study. We will use this project to illustrate our points in this paper.

In this paper we present some of the challenges for building human-inspired interactive systems, pointing out the complexity of the task that is partly determined by the nature and the structure of communicative processes, the many variables

that play a role, and the many forms that communication can take. But most importantly, the complexity of the task arises because of methological restrictions.

In Sections 2 and 3 we will concentrate on what it means to develop human computing systems that are able to converse with humans. Theories of conversation provide a model of conversations and the terms that are used to name the important distinctions, structures and phenomena in conversations. Based on models such as these the parameters that make up an information state in a dialogue system are defined. Similarly, the theory provides us with the labels to use in annotation schemes for the description of actions and events in multimodal corpora. This is what we will focus on next. We discuss, how reliability analysis of the annotations play a central role in the evaluation of human computing systems. Without it a corpus is not of much use. Reliability analysis forms the interface between the quantitative measures of the phenomena and the qualitative and subjective measures of the same phenomena. They form the contract between systems designer and end user of the system. We illustrate these issues by discussing our analysis of feedback in the AMI dialogues in Section 5. More about the AMI corpus and the place of the scenario based meeting recordings in the methodology can be found in [McCowan et al., 2005] and on the website: http://www.amiproject.org.

The aim of this paper is to understand the *possibilities* and *boundaries* of human computing technology. Human computing, we believe, means building interactive systems that have the same capabilities that humans have in communication. We provide an analysis of what it means to "engineer natural interaction". This analyses may inspire new approaches to design such systems, and they may inspire new directions of research in human-human and human-computer interaction.

2 Engineering Natural Interaction: Preliminaries

Every student after building his first natural language dialog system knows the disappointment when he demonstrated his system to his teacher and the first sentence that was typed in or spoken did not work. "Oh, yeah, sorry, if you want to say something like that, you have to say ...". Sometimes followed by a promising: "I can easily build that one in." A teacher feeling sympathy for the students attempt will respond with: "Well, you can't have it all. It's the basic idea that counts. How does your system work? How is it build up? And, how did you design it?"

Montague considered English a formal language ([Montague, 1970]) and in natural language processing this kind of view on the equation between natural and formal is the basis for developing theories, algorithms and applications. If we put off our formalist glasses we see that people don't speak grammatical sentences; fragmented speech, restarts, incomplete sentences, hesitations, insertions of seemingly meaningless sounds, are more the rule than the exception. It is important to remember that the formalisations in NLP, inspired heavily in the first years by Chomsky's view on language as a system of rules, are based on an

idealisation as well as a simplification. The early Chomskian view entailed abstracting away from performance to competence and from the actual language user to an "ideal" speaker. It focussed on formal aspects such as syntax and ignored language use ([Chomsky, 1965]).

Computational engineering, by nature, always follows a similar path of abstraction and formalisation. NLP systems rely on dictionaries and grammars, that try to capture the rules and regularities in natural language that make communication possible. It is this "static side" of language that we can build natural language technology on. This can be complemented by *data-driven*, *statistical* approaches that try to bridge the gap between what users do and the functioning of our natural language processing system. But natural language is not something fixed, something that is understandable in the form of a formal system. Communication is build on conventions and regularities but at the same time it is essentially a way to negotiate meaning and to establish new conventions for interaction. Technology therefore will always fall short.

In technology, the conventions are present in two ways. They are fixed, by the engineer (just as dictionaries and grammars fix a language in a particular state) and they are communicated to the user of the system in the form of a user manual which is essentially a *contract* between the user and the technology. The contract specifies the interaction with the user; what it affords, what the semantics of the expressions of the user in communication with the systems is, in terms of the effects the systems brings about in response to this. The contract further specifies how the user should interpret the systems outputs. In short, the contract defines the user interface with the system in the practice of use. When the user asks a question in a certain way, the contract says that the system considers this as a question and that it will respond to it in a certain way. In systems that aspire to natural interaction the forms of interactions are made to look as much as possible like the forms of interactions the user already knows from the experience of interacting with natural systems.

We will repeatedly use the term contract to identify a set of interactions between the phenomena as they 'naturally' occur and the scientific study of them, between the scientist and the engineer that uses the knowledge to construct systems, and between the engineer and the user of the system. If all works out well engineered interactions will become second nature.

3 The Nature of Communication

In this section we will outline how human-human communication works and what it would mean to build machines that can make sense of what is going on when humans communicate and what it would involve for a machine to participate in a conversation. The phenomena and processes that make up conversations are presented in the form of variables in so-called information states in dialogue systems and enter in the guise of labels used in annotations of recordings of dialogue, as we will see later.

To start the discussion, imagine a group of people having a conversation. Look around you or turn on the television and think about what you are seeing.

Participants in face-to-face conversations move their heads, their bodies, their hands, their lips, their eyes. They produce sounds among which there are sounds that one might identify as expressions of some natural language. When we look at the whole collection of behaviours we segment them in all kinds of ways and label them with a variety of categories: a left eye-brow is lowered, a cup is grasped, a person is speaking, another is listening, somebody asks a question, the speaker is embarrassed, people shake hands, they interrupt each other, they threaten each other, they enjoy themselves. Often the same behaviour can be classified in many ways; we can describe the action in physical ways (mouth corner pulls up), or categorize the behaviour using a common word for the action (smile), characterize the function (she relaxes) or the effect (she made him think she likes him).

There are a couple of things we do when we are involved in a conversation. The behaviours that we see or hear displayed by our conversational partners are not only identified and classified in all kinds of ways. We also consider what caused them, what function they served, what intention (if any) was behind them. What we call communicative behaviours are typically those that were intended to be recognized as serving the intention they were intended to serve. One of these intentions is that the action would be recognized as a communicative behaviour but typically there are many others besides this one. A way to group theintentions is given by the list of levels that [Clark, 1996] distinguishes. Being recognized as a communicative behaviour is, one could say, the second one on this list: the level at which a person producing some communicative behaviour intends it to be identified as a signal (a meaningful act intended to be taken as meaningful by the receiver).

1. Joint[A executes behavior t for B to perceive; B attends perceptually to behavior t from A]
2. Joint[A presents signal s to B, B identifies signal s from A]
3. Joint[A signals to B that p, B recognizes that A means that p]
4. Joint[A proposes a joint project to B, B takes up the joint project]

This list of levels at which a communicative behaviour functions shows both the actions performed by the communicator (the speaker, in case of speech) as well as the corresponding action by the addressee (the hearer, or whatever). The notion of communication as a joint action is one of the ways to view an inherent feature of communication: the fact that behaviours, actions and intentions of participants in a conversation are tightly coupled and directed at each other. This fact leads [Schegloff, 1982] to view a conversation as an interactional achievement. This way of putting it stresses another point: that the way a conversation evolves, the way actions of a participant unfold are contingent on the actions and responses of the other participants. Clearly if A executes some behaviour for B to perceive, but notices that B is not attending he will have to take measures in order to ensure the success of the communication. As [Kendon, 1967] remarks: participants function in two modes (at the same time): an expressing and a monitoring mode.

Monitoring the other participants is an important part in the process of conversation as a person who makes a contribution should check whether his action

has succeeded and this is dependent on the changes effected in the state of the other person to which the contribution was directed. [Clark and Schaefer, 1989] consider what it takes to make a contribution to the discourse to consist of two participatory actions by the contributor and the addressee: the contributor presents his utterance (presentation phase) and the addressee provides evidence that he has understood the utterance (acceptance phase). The following main types of providing evidence are distinguished in [Clark and Schaefer, 1989]:

1. Continued attention by the addressee
2. Initiation of the relevant next contribution
3. Acknowledgement
4. Demonstration
5. Display

The first type shows that behaviours that are produced to monitor and to attend can by themselves be indicative and expressive. The most typical example of this is the pattern described by [Goodwin, 1981] who notices in the conversations he is studying that a speaker as he makes a contribution will make sure that the addressee is looking at him and if not will pause till the addressee does so. By initiating a next contribution the addressee shows that he has at least understood the contribution by the previous speaker as an invitation to make some contribution. Of course, this next contribution may make it clear that the addressee did not understand the utterance as intended at all. Acknowledgments consist typically of nods and backchannels such as 'uh huh' and 'yeah'. If an addressee performs or starts to perform the action that the contributor was inviting him to perform, than this is a typical demonstration of understanding. Finally, a display of understanding is taken to be a case where the addressee displays verbatim all or part of A's presentation.

One of the important factors that determines which kind of behavior is displayed as evidence of understanding, is the precise setting and the task that has to be performed. In face-to-face conversations continued attention is often shown through gaze or nonverbal acknowledgments. In contexts as described in [Nakano et al., 2003] and [Kraut et al., 2003] on the other hand, where one person instructs another on physical tasks, the right or wrong execution of the task provides information on understanding.

An important lesson to draw from these and many other examples is that contexts and setting (physical context, task context) determine to a large extent the precise kinds of behaviours that are being performed, whereas the underlying functions that require behaviours of one kind or another remain the same.

The behaviours that recipients of a communicative behaviour engage in as a response to show the acceptance (or failure of acceptance) constitutes a communicative behaviour in its own right and will be responded to by the other participants. A typical case is where a speaker sees that a listener is paying continued attention to what the speaker is saying to which the speaker responds by continuing to speak.

There are many cases where observers and participants in a conversation fail to interpret signs and signals or whether acceptance and understanding is only partial. The perception of the behaviours displayed may be faulty. The behaviour may not be recognized as bearing a meaning or may be interpreted different from intended. And of course: understanding what someone wants to achieve does not necessarily lead to agreement and acceptance of the joint project that is being proposed. Typically, the acceptance may be partial in other ways as well. For instance, in the case where one is told something with the purpose to change one's beliefs, one may find the message implausible, unconvincing, untrustworthy, or highly plausible but with some scepticism remaining. So besides the fact that acceptance proceeds on different levels it can also proceed to different degrees. It depends on the context and the goals of the interlocutors what degree is acceptable. For instance, a belief may have been communicated by a speaker for the purpose of informing the addresse about the speaker's belief set or it may have been communicated to actually convince the addressee of the truth of this belief.

Having the addressee update his beliefs (as is the case for the class of speech acts called *assertives* by Searle) or showing how one feels (as with the class of acts called *expressives*), are not the only kind of purpose that a communicative act may serve and the kind of participatory behaviours that interlocutors display in response may differ accordingly. Note also that many conversations involve more than two participants. Communicative behaviours may be directed at multiple participants at the same time with the aim to have a different effect on each of them.

Not all the behaviours displayed in a conversation are intentionally produced to communicate. Eye-blinking or breathing are some of the stereotypical behaviours that go on automatically, unconsciously. Most of the time they also go by unnoticed. Nevertheless they may work as natural signs, providing information about the mental and physical state of a participant. Yawning and sneezing are behaviours that are mostly produced without the intention to communicate as such, but they do have an important impact on the conversation.

Another typical case of unintentional information production is the case of leakage which is often discussed in the context of non-verbal communication when particular behaviours that are difficult to control by a person provide information contrary to the intentionally produced utterances. It is foundational for human communication that we are able to distinguish the semiotic status of a behaviour. If we see someone raise his hand we can understand this behavior in three different ways ([Buytendijk, 1964]):

1. as an *expression*, of anger or condemnation, for example
2. as a part of some *action*, the killing of a gnat
3. as a *representative gesture*, a greeting for example.

We therefore have to consider the whole situation and how it develops in time. The distinction between an act and an expression is that an act is a movement that is directed towards some specific endpoint, its goal, whereas an expression is a movement which is an image, in which a meaning becomes visible. As far as a behavior is an expression it shows some inner state of the subject, affect or mood; as far as it is an act, it is a movement that is performed to establish some outside

state of affairs. Eye blinking, for example, can be an act (to remove something from the eye) or an expression of nervosity, and it can also be a representative gesture. In the latter case the act has become a sign (a wink) and the movement looses its primary function. In gesturing, such as speech and writing the relation between the movements and their senses, the meanings they intentionally refer to is indirect, whereas in the expressions the inner state is immediately revealed. If someone stamps his feet on the ground this can be an expression of anger as well as a gesture to indicate that one is angry, or both. Machines only act, and even that only in a metaphorical sense; they don't have an inner state that is revealed in expressive behavior. Animals show expressions and can act as well, but they don't make gestures. It should be noted that it is sometimes hard to tell whether a certain behavior is expression or act. Almost all acts and all gestures (such as speaking) have the character of an *expressive* behavior[1] also, through the *way it is performed*.

Being able to identify the physical characteristics of a behaviour is a necessary prerequisite for interpretation but the challenge of interpretation resides in knowledge about the situation in which the behaviour takes place.

The interpretation of (communicative and other) behaviours involves a search for the determinants that caused the behaviour. Intentions and rational goals that are obviously important determinants, but there are many other kinds of determinants that play a role in conversation. Expressing or hiding how one feels, the need for affiliation and contact, social obligations and commitments that need to be taken care of, are some of the concerns that [Goffman, 1976] would classify under the ritual concerns. One way to classify the various motives for communicative behaviors is by the following needs.

1. The need to get something done: business, tasks, goals.
2. The need to communicate and build up rapport.
3. The need to express oneself.
4. The need to make conversations go smoothly.

Besides the task goals, the interpersonal and the expressive side, there is also a concern with the way the communication proceeds as such: metacommunicative goals that may involve concerns with turn-taking, channels of communication etcetera. The system constraints that [Goffman, 1976] lists provide a good indication of what these meta-communicative actions involve or what is necessary for smooth communication.

1. A two-way capability for transceiving readily interpretable messages
2. Back-channel feedback capabilities "for informing on reception while it is occurring"

[1] Studies of behavioral signals in affective computing research often consider only the expressive function, ignoring the fact that most signals (such as facial expressions) in dialogue constitute discourse-oriented actions, i.e. linguistic elements of a message instead of "spillovers" of emotion processes ([Bavelas and Chovil, 1997]).

3. Contact signals (sigalling the search for an open channels, the availability of channel, the closing of a channel, etcetera)
4. Turnover signals (turn-taking)
5. Preemption signals: "means of inducing a rerun, holding off channel requests, interrupting a talker in progress"
6. Framing capabilities: indicating that a particular utterance is ironic, meant jokingly, "quoted", etcetera.
7. Norms obliging respondents to reply honestly in the manner of Grice's conversational maxim.
8. Constraints regarding nonparticipants which should not eavesdrop or make competing noise.

Besides such system constraints that tell how individuals ought to handle themselves to ensure smooth interaction, an additional set of constraints can be identified "regarding how each individual ought to handle himself with respect to each of the others, so that he not discredit his own tacit claim to good character or the tacit claim of the others that they are persons of social worth whose various forms of territoriality are to be respected." These are ritual contingencies that need to be taken into account and that may also take up a couple of exchanges. Also back-channel expressions may let a speaker know whether or not what he is conveying is taken to be socially acceptable besides signalling understanding. Because conversations are joint actions, the desire of one person to communicate must be matched with the will of the other to participate in the conversation as well. Conversation thus involves a complex structure of negotiating rights and obligations regulated by norms and social conventions.

Modeling conversational action for automatic processing (both analysis and generation) requires not just the modeling of how rational actions are performed through language, how the mechanics of language and conversations work, but also of the personal, interpersonal concerns, emotions and attitudes play a role.

In the next section we look at the way these aspects of conversation are taken into account in current systems that analyse and reproduce natural dialogue.

4 Ways to Engineer Natural Interaction

The main challenge for ambient intelligent systems is to be able to determine what is on the mind of a person, inferring this from the behavior displayed in the particular context, prior knowledge about the person and about the behavior of humans in general. Although conversations are a particular type of action and other interactions between humans and the ambient technology need not be conceived of as following this model in every detail (for instance, by having embodied conversational agents or humanoid robots all over the place) they are well suited to illustrate the issues of natural interaction as conversations display the full gamut of processes and modes of interaction. The structures and patterns in interaction, the processes as they have been identified above, by presenting

a view on conversation and the way it is modelled in theoretical and practical frameworks, provide an overview of all the things to take into account when modeling and implementing human-system interaction in a natural way. As we said before, though, in different contexts the precise forms the contributions take and how they are organised will differ. It is therefore important to use methods that can deal with this contextual dependence. The development of spoken dialogue systems often proceeds by starting with collecting Wizard of Oz data and taking the communicative behaviors that people deploy in these types of interactions as representative for future interactions with the system.

We will now point out two kinds of research areas which formalise the phenomena that make up conversations. One is the practice of building dialogue systems, where the phenomena turn up as variables in an information state and the second is the study of algorithms for automatic analysis of human-human interaction, where the phenomena turn up as labels to describe the data.

Consider again the four levels described by [Clark, 1996]. If we think of a system that is engaged in conversations in a similar way the system should: 1) be able to execute particular behaviors, 2) that count as signals 3) with an intended meaning 4) in an effort to propose a joint project, and 5) be able to perceive behaviors from others, 6) identifying them as signals 7) and recognizing their meanings 8) so as to figure out and take into consideration the project that is being proposed.

Algorithms developed for specific applications may focus on one or more of these aspects. Current dialogue systems offer an example of the way in which the elements and processes that make up a conversation can be conceived of in terms of data structures and algorithms that keep track of the most important variables. In many dialogue systems what goes on in a conversation is captured in an information state that is updated as the conversation proceeds, often with a stack of states capturing the history of the conversation. One of the more complex instances of such a state is presented in [Traum and Rickel, 2002]. A multiple layer approach is taken in this paper towards modelling and managing the complexities involved in multi-party multi-modal interactive systems, "including who is accessible for conversation, paying attention, involved in a conversation, as well as turn-taking, initiative, grounding, and higher level dialogue functions".

The central Information State, a store of information, that is updated by functions, that are the interpretations of the Inputs received by the Interpreters. The Generator module uses to updated Information State to decide for the actions to be performed and generated.

The following layers are distinguished in [Traum and Rickel, 2002]. This list of layers and parameters bears close resemblance to the list of system constraints we have presented in the previous section.

1. Contact layer: whether and how individuals can communicate: the modalities that can be used and the media that can be used. (*make-contact, break-contact*)
2. Attention layer: the focus of attention of each of the participants (*give attention, withdraw attention, request attention, release attention, direct attention*)

3. Conversation layer: model of the various dialogue episodes going on throughout the interactions (there may be several conversations going on in parallel)
   (a) Participants: active speakers, addressees, overheareres, etc.
   (b) Turn: the participant with "the right to communicate" using the primary channel (take-turn, request-turn, release-turn, hold-turn, assign-turn)
   (c) Grounding: how is information added to the common ground
   (d) Initative: the person who is controlling the contributions: take-initiative, hold-initiative, release-initiative
   (e) Topic: start-topic and end-topic
   (f) Rhetorical connections between content units
4. Social commitments
5. Negotiation layer

Contributions to dialogue will typically perform several functions at the same time and will thus be multiply determined. The central research question is whether we can determine how the many behaviors determine the update of the various functions and how we can use this knowledge to analyse human behavior and generate appropriate responses.

The definition of the layers and the variables in the information state of a dialogue system is an instance of the formalisation that turns natural processes into a formal architecture. The same kind of objectivation of theories and assumptions about conversation takes place when corpora of human-human interaction are collected and annotated. The specification of the coding schemes puts down how terms and concepts will be applied to specific instances of real data. We have been working in particular on the AMI data, as mentioned in the introduction. In the case of the AMI corpus, this resulted in the following levels:

1. Named Entitities
2. Dialogue Acts
3. Meeting Actions
4. Emotion and Mental state
5. Topic Segmentation
6. Text Transcription
7. Individual Actions
8. Argumentation Structure
9. Focus of Attention
10. Person Location
11. Various kinds of summaries

There are several important research questions for human computing in these contexts that are typical for data-driven, corpus-based research. First, how can we derive metadata automatically from the raw signals; for instance, automatic transcripts using speech recognition, or descriptions of facial expressions like action units from the video using computer vision techniques). Second, using the hand-made and (semi-)automatically derived metadata to infer further information about what happened in the meetings. We will illustrate this second

question below. Third, the corpus and its metadata can be used to derive certain statistics about behaviors as they occur in the corpus that can be used to test certain hypotheses about human behavior or as input for the construction of artificial entities that need to respond to or display similar behavior. Testing the assumptions about conversations derived from theories, does not only happen after the annotation has been performed but also at the development stages of the annotation scheme and the initial tests.

The collection of data and metadata can serve different purposes depending on the context. For instance, in one case information that is present may be used in the algorithm as one of the parameters on which the algorithms bases itself to further classify and label. In other cases, the metadata derived from manual annotations may be used as the ground truth with which automatically derived data of the same kind is compared for evaluation purposes.

An important methodological issue in the collection of corpora and the construction of the annotations is the specification of the labels to use for description and the definition of their use: to what kinds of objects do they apply and how should an annotator decide what counts as what. This issue is addressed mostly by using an iterative approach, where initial drafts of schemes are tested on subsets of the data by several annotators to find out the fit between theory and data and the precision of the specification by measuring the intersubjective agreement. We illustrate these steps in the remainder of this section.

We build three types of models. The first type comprises the *qualitative* models, in which we *describe* what happens in meetings using terms from the various scientific vocabularies to express the important concepts, ideas, the phenomena and processes that we observe in meetings. The second type consists of the *quantitative* models which can be rule-based or statistical. The third type contais the *computational* models, software implementation and their implementations. In each of these models the words and notions always keep referring to the intuitive semantics of the primary concepts that we know from our practical experience with meetings.

What we will point out in the next sections is how the various models are connected. The first model provides data from which to derive the second kind of models. The second kind is used to develop the computational models. Both of the latter kinds, can provide us with insights that make us change our theories and models of the data, which leads to an update of our theories and possibly our data annotations in an incremental way.

5 From Data Analysis to System Integration

In our work on building affective dialogue systems or other applications in which human communication is processed automatically, corpus collection, corpus annotation, and reliability analysis of the annotation procedures play a central role. In this section we present this *methodology* and how the various steps fit in the proces of designing and evaluating a human computing system. We will show how this method is based on a number of feedback loops. To illustrate this we

present some details about our studies on *feedback* behavior of listeners in face to face meetings where the analysis of the data leads us to rethink the notions that we started out from.

5.1 The Method

The method that is usually followed basically consists of steps that are motivated by the specific application that one has in mind. This can be to make a system that recognizes facial expressions of affective states of learners in a face-to-face tutoring situation, a system that recognizes certain backchannels and turn-taking behavior in face-to-face conversations, or a system that has to generate rich expressive speech for a virtual story teller. In all of these cases, what we are aiming at is to model natural human behavior, to implement it so that our system behaves as humans would do in similar situations. The steps we take are the following:

1. data-collection
2. data-analysis and modeling
3. model-implementation
4. system evaluation
5. reconsideration

Data-collection can either be done in natural situations, but often happens in controlled situations, similar to experimental physics. A major point of concern here is that the situation is such that we can reasonable expect that the results of our analysis can be transferred to the situation in which we want to apply the model. Ecological validity is very important in this kind of research. Based, in part, on intuition and state of the art theoretical findings, we design annotation schemes in which we define labels for the relevant phenomena of interest in our data.

The annotation procedures are used by human annotators to produce hand-annotations. Other features are computed automatically, such as for example the F0-contour of speech signals, or the movements of facial units, or the words that occur in the realisation of some dialogue act. Hand-annotation by human annotators is essential here since not all features can be automatically identified.

Reliability analysis is then carried out on the annotations to see whether different annotators agree sufficiently in the way they labeled the phenomena. If the measure of agreement is too low we can either throw away that part of the annotations that show low agreement or we will redesign the annotation proces. Reliability analysis is an essential step because we need a reliable relation between the features that quantitatively describe a phenomenon and the qualitative label that is assigned to it. Only then, as engineers can we offer the users a contract that is the basis for the application of the system we build and the way the user interprets the outcomes or expressions that the system produces.

The models that we build based on the data analysis can either be statistical or analytical. If we use automatic classifiers trained on the annotated data we hope

that the classifier performs accurately on unseen data. Note that a high inter-annotator agreement and a high accuracy score of the classifiers is an indication that we have succeeded to model the phenomenon in an accurate way. Evaluation with users, other subjects than annotators, should see whether the outcomes of the machine conforms the way the subjects assess and label the outcomes.

For all practical purposes reliable annotations and a good classification method, provide us with a contract that guarentees the soundness of our design and a manual for the potential users, as long as they use it in a context that satisfies the conditions of our experimental situation.

Still, since we are dealing with data from a limited number of humans annotated by a limited number of annotators our evaluations can only give us statistical outcomes: the best we can offer is saying something like "in 95% of the cases what the facial expressions of the ECA in this type of situation will show is a grin". It is not possible to pin this down to a statement about this unique situation. Moreover, the behavior that is shown by the system will be some statistical mean, representing the "average" behavior of the subjects that happened to act in the data collection proces.

5.2 Feedback in Conversations

Our analysis of feedback or bachchannels[2] in the AMI corpus provides us with a good example of how the various steps in the method outlined before relate to each other: how one level feeds into another and back.

The starting point of our research was technological but accompanied by other research questions as well. In multi-party interactions it is not always obvious who is addressed by an utterance of the speaker. Being able to detect it automatically is an important challenge. As we have seen in Section 3, conversational acts are joint actions where actions of speakers are complemented by actions of listeners. The recipients of communicative behaviours will typically display certain behaviours that provide feedback to the speakers. From this it should follow that if we can recognize this feedback behaviour, we may use this to identify a potential addressee of the speaker action. Clearly, someone who provides feedback, felt addressed in some way or another. Similarly, certain actions of speakers

[2] The term backchannel was coined by Yngve (1970) and is derived from the notion of a "back channel" through which the listener sends the speaker short messages, such as "yes" and "uh-huh", that are not a bid for the floor. Which types of utterances can be considered backchannel activity is often debated. The very short messages like "mmm," "yeah," "right," -which are common in English- clearly qualify because they add a great deal to the quality of the interaction without really adding meaning to the conversation. However, Yngve also considers questions such as, "You've started writing it, then, ... your dissertation?" and short comments such as, "Oh, I can believe it," to be backchannel utterances. Duncan [Duncan and Niederehe, 1974] added other types of utterances to the list, such as sentence completions, requests for clarification, and brief restatements, since their purpose is not actually to claim a turn, but to provide the speaker with needed feedback.

may provide us with information on whom he is addressing as well. It has been pointed out (for instance by [Goodwin, 1981]) that speakers may indicate whom they are addressing by looking at them at certain points in the utterance. The combination of feedback and gaze cues also leads one to an hypothesis about their co-occurrence. Could it be a regular feature of conversation that feedback of listeners occurs at positions where the speakers gaze at them? In those cases, listeners know that they are being addressed and that speakers are attending to them and can perceive the feedback. Information about statistics such as these[3] can also be used in the design of our conversational agents. In particular, we have been looking at the implementation of agents that can provide appropriate feedback and this kind of information would help in the timing of the feedback ([Heylen, 2007]).

The AMI corpus consists of more than 100 hours of video and audio recordings of four person meetings. We already mentioned the annotation layers with which the data was enrichted. Several of these are relevant to answer this question. The hand-coded dialogue acts contain several labels for feedback acts and other relevant information: on the relations between different utterances and on addressing. Information on the focus of attention, to whom or what somebody is looking, is also present in the AMI corpus. The corpus and the annotations thus appear to be ideal to answer our questions. However, for several reasons the solution was not as straightforward as may appear.

The AMI dialogue act annotation manual distinguishes three types of feedbacks: Backchannel, Assess and Comment-about-understanding. The Backchannel class largely conforms to Yngve's notion of backchannel and is used for the functions of contact. Assess is used for the attitudinal reactions, where the speaker expresses his stance towards what is said, either acceptance or rejection. Comments about understanding are used for explicit signals of understanding or non-understanding.

In addition to dialogue acts the coding scheme specifies that certain relations between dialogue acts should be annotated. Relations are annotated between two dialogue acts (a later source act and an earlier target act) or between a dialogue act (the source of the relation) and some other action, in which case the target is not specified. In the AMI scheme, relations are a more general concept than the adjacency pairs from the Conversational Analysis literature, like question-answer. Relation have one of four types: positive, negative, partial and uncertain, indicating that the source expresses a positive, negative, partially positive or uncertain stance of the speaker towards the contents of the target of the related pair. For example: a "yes"-answer to a question is an inform act that is the source of a positive relation with the question act, which is the target of the relation. A dialogue act that assesses some action that is not a dialogue act, will be coded as the source of a relation that has no (dialogue act as) target.

Since Backchannels were assumed to be always in response to what the main speaker at that point is saying, annotators did not enter them in a relation with another dialogue act, assuming that this could be detected automatically. But

[3] See also [Heylen, 2006] and [Poppe et al., 2007].

in order to check the relation between the gaze target of the speaker and the one who gives the feedback, we need to find for each occurrence of a backchannel act the related dialogue act that the backchannelor responds to, as well as the speaker of this related dialogue act. As the AMI dialogue act annotation does not contain the annotation of the relation between backchannel acts and this dialogue act (or turn), we have to define a new challenge: define an algorithm that decides which utterance a back-channel is related to.

We implemented and tested a procedure for doing this. We have validated our method for finding the related dialogue act (measured by recall and precision) and selected some parameter values for the time between act and backchannel act that gave us the best performance. The results can be found in the following table.

	correct	incorrect	uncertain	total
found	77	3	3	83
not unique	44		17	61
total	121	3	20	144

Fig. 1. Results of the method for finding the related dialogue act that a backchannel responds to

For 83 out of the total of 144 backchannel events the procedure reported to have found a unique related dialogue act. Of these 77 was the correct one, 3 were incorrect and in 3 cases the answer was questionable. In these cases it was actually not clear what the related utterance is. Of the 61 cases in which the procedure reported that no unique related act was found, there were 44 cases in which there was a unique related dialogue act but the algorithm failed to identify it. In 17 cases it was unclear also from manual inspection to identify a related dialogue that the backchannelor responded to. If we leave out these uncertain instances we end up with 124 cases, and 77 correct and 3 incorrect answers, hence the method has a recall of 77/124 (62%) and a precision of 77/80 (96%).

When we checked the outcome of the procedure the major causes for not finding a *unique* dialogue act were the following.

1. There is simultaneous speech of multiple speakers. This occurs in animated discussions, where speakers sometimes express their ideas in cooperation. The backchannelor is responding to the idea expressed not to one speaker.
2. A particular situation arises when a speaker pauses then continues and right after the continuation the backchannelor is reacting on the part before the short pause. The method will not find the correct act because it seems to respond to the continuation and not the previous part. This is a serious case because continuer signals often occur in the middle of speaker turns where short pauses or segment boundaries occur.

It is also interesting to look at the case where the relation between backchannel and the related utterance was also unclear in the manual annotation.

The following cases were found where the situation was indeed unclear to identify a related speaker and act. These are cases

1. with simultaneous cooperative talk
2. with the absence of a dialogue act (in particular "backchannels" such as "Okay" were used mostly as a closing signal)
3. where back-channels appeared to be instances of self talk ("Mmm", "yeah") and not directed to a particular other contribution.

The analysis we performed shows how initial assumptions that inform the annotation of data might need revision.

Next we looked at the relation between speaker gaze and the occurrence of feedback. For each of the 13 meetings we computed for each pair of participants (X, Y) the length of time that X looks at Y while X is speaking and we computed the length of time X looks at Y when X performs a dialogue act and Y responds with a backchannel act. Analysis of these pairs of values shows that in a situation where someone performs a backchannel the speaker looks significantly more at the backchannelor than the speaker looks at the same person in general when the speaker is performing a dialogue act ($t = 8.66$, $df = 101$, $p < 0.0001$). The mean values are 0.33 and 0.16. This confirmed our hypothesis.

This example shows a number of issues related to the methodology that we and many others follow in human computing research. We collect and describe data for our data-driven methods, building on theories of human communication. The theories and insights that lead to the development of an annotation scheme may not be full-proof. The cracks in the theory or the cases that are not covered can be detected by the cycle of research that is exemplified here. In this way, engineering is based on knowledge but also the basis of knowledge. In our case, it lead us further to rethink the theories of participation and action in conversations that formed the basis for our initial annotation framework. However, this does not mean that we have now reached the ultimate theory of communication. This leads us to the conclusion.

6 Human Computing

One can view human computing technology as the current state of the historical development of technology, that aims at simulating human interaction as an aspect of human behavior as such. It shows the boundaries of the scientific methodology, inherited from mathematics, physics and technology, and the way we conceptualize nature and human behavior, according to the principles of this world view. What is central in this development is a principle conflict between on the one hand the natural openness of natural language and human behavior, in which the human mind freely assigns through his practice and in communication with the world and others new means of signifying, and new processes of control. Technology itself is the phenomenon where this creativity shows and this means that technological development will always be essentially an incomplete objectification of human creative, sense giving, behavior. We have pointed out some

of the consequences that this has for the design of human computing systems. We have put emphasis on the role of the *contract* between user and technology, its essential role for the working of technology in practice. The contract is the issue that has to be evaluated, if we evaluate our technical systems in practice, because it forms the interface between the qualitative measures of the user and the quantitative measures that function in the design of the system.

The identity of the "virtual human" that technology brings forward, i.e. the ambient intelligence that is our technological communication partner, is essentially a mathematical identity, either in the form of a statistical model or of an analytical model. When we construct this technology we have the obligation to "show" that it satisfies the contract that is implicit in its design. The reliability of the data-analysis is essential for the contract being possible and for the ecological validity of the experiments that underly the models that are implemented. The contract between users and system designer make up the context for the services that the system offers the user in using the technology in his practice.

The great challenge for human computing is in further clarification of the central concepts that play a role in the embedding of technology in interaction with man; concepts like service, what technology affords, and context, and related notions as context-awareness.

Acknowledgements. This work is supported by the European IST Programme Project FP6-0027787. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[Bavelas and Chovil, 1997] Bavelas, J. and Chovil, N. (1997). Faces in dialogue. In Russell, J. and Fernandez-Dols, J.-M., editors, *The psychology of facial expression*, pages 334–346. Cambridge University Press, Cambridge.

[Buytendijk, 1964] Buytendijk, F. (1964). *Algemene theorie der menselijke houding en beweging*. Het Spectrum, eight printing 1976, (in Dutch).

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of a Theory of Syntax*. MIT Press, Cambridge Massachussets.

[Clark, 1996] Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge.

[Clark and Schaefer, 1989] Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259294.

[Duncan and Niederehe, 1974] Duncan, S. and Niederehe, G. (1974). On signalling that its your turn to speak. *Journal of Experimental Social Psychology*, 10:234–47.

[Goffman, 1976] Goffman, E. (1976). Replies and responses. *Language in Society*, 5(3):2257–313.

[Goodwin, 1981] Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York.

[Heylen, 2006] Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267.

[Heylen, 2007] Heylen, D. (2007). Multimodal backchannel generation for conversational agents. In van der Sluis, I., Theune, M., Reiter, E., and Krahmer, E., editors, *Workshop on Multimodal Output Generation*, pages 81–91, University of Twente. CTIT.

[Heylen et al., 2005] Heylen, D., Ghijsen, M., Nijholt, A., and Akker op den, H. (2005). Facial signs of affect during tutoring sessions. In Tao, J., Tan, T., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction - First International Conference, ACII 2005*, pages 24–31. Springer-Verlag. ISBN=3-540-29621-2.

[Heylen et al., 2004] Heylen, D., Vissers, M., Akker op den, H., and Nijholt, A. (2004). Affective feedback in a tutoring system for procedural tasks. In André, E., Dybkjr, L., Minker, W., and Heisterkamp, P., editors, *ISCA Workshop on Affective Dialogue Systems, Kloster Irsee, Germany*, pages 244–252, Berlin Heidelberg New York. Springer-Verlag. ISBN=3-540-22143-3.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

[Kraut et al., 2003] Kraut, R., Fussell, S., and Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction*, 18:13–49.

[McCowan et al., 2005] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., M.Kronenthal, Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. In *Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research*.

[Montague, 1970] Montague, R. (1970). English as a formal language. In Visentini, B., editor, *Linguaggi nella società e nella tecnica*, pages 189–224. Edizioni di Comunità, Milan.

[Nakano et al., 2003] Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 553–561. ACL.

[Pantic et al., 2007] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2007). Machine understanding of human behavior. In *Proceedings AI for Human Computing (AI4HC'07). Workshop at IJCAI 2007*, pages 13–24, Hyderabad, India.

[Poppe et al., 2007] Poppe, R., Rienks, R., and Heylen, D. (2007). Accuracy of head orientation perception in triadic situations: Experiment in a virtual environment. *Perception, to appear*.

[Schegloff, 1982] Schegloff, E. A. (1982). Discourse as interactional achievement: Some uses of "uh huh" and other things that come between sentences. In Tannen, D., editor, *Analyzing discourse, text, and talk*, pages 71–93. Georgetown University Press, Washington, DC.

[Traum and Rickel, 2002] Traum, D. and Rickel, J. (2002). Embodied agents for multiparty dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773.

Evaluating the Future of HCI: Challenges for the Evaluation of Emerging Applications

Ronald Poppe, Rutger Rienks, and Betsy van Dijk

University of Twente, Human Media Interaction Group
Enschede, The Netherlands
{poppe,rienks,bvdijk}@ewi.utwente.nl

Abstract. Current evaluation methods are inappropriate for emerging HCI applications. In this paper, we give three examples of these applications and show that traditional evaluation methods fail. We identify trends in HCI development and discuss the issues that arise with evaluation. We aim at achieving increased awareness that evaluation too has to evolve in order to support the emerging trends in HCI systems.

1 Introduction

The field of Human-Computer Interaction (HCI) is concerned with the research into, and design and implementation of systems that allow human users to interact with them. Traditionally, the goal of HCI systems is to aid human users in performing an explicit or implicit task. Currently, there is a shift in emphasis towards interfaces that are not task-oriented but rather focused on the user's experience. More subjective factors such as the beauty, surprise, diversion or intimacy of a system are important [1; 2].

A vast body of literature deals with evaluation of traditional HCI systems. These evaluation methods are widely used. However, given the new directions of HCI, it is unlikely that these evaluation methods are appropriate.

Recently, the term Human Computing (HC) has been introduced. In this paper, we discuss Human Computing and the differences with Human Computer Interaction in Section 2. We outline new trends in HCI systems in Section 3. Section 4 presents three examples that illustrate the need for new evaluation methods. In Section 5, we discuss common evaluation methods, argue why these are inappropriate and identify challenges for evaluation of emerging HCI systems.

2 Human Computing and HCI

There is clearly an overlap between Human Computing (or, alternatively, Human-Centered Computing (HCC) [3]) and Human-Computer Interaction. Both deal with humans who interact with computers or machines. The role of the user is central in both paradigms, but there is a difference in the extent. In Human Computing, the user and its contexts are not only observed, but the user's intentions and motives are estimated from the observed behavior. In turn, the system is to display behavior that informs the user about the intentions and motives of the system. For both the observation and

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 234–250, 2007.

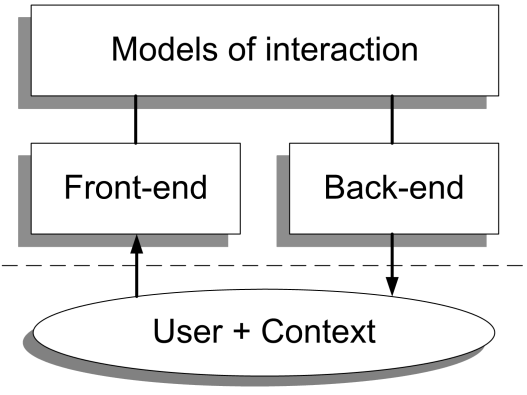


Fig. 1. Schematic model of Human Computing

presentation of intentions, models of interaction are required. For successful, natural, interaction, these models should be as close to human-human interaction models as possible. Due to their importance, these interaction models are explicitly part of the Human Computing paradigm. The concept is illustrated in Figure 1.

In Figure 1, there are clearly three distinct parts. The front-end deals with observing the user and its context. Aspects like the activities, affective state, but also the context such as the environment and other users are taken into account. Pantic *et al.* [4] discuss the front-end, with a focus on recognition of the user's affective state. The back-end is concerned with the presentation of information to the user and the control of actuators in the environment of the user. Different from HCI are the models of interaction, that are used both in the front-end to understand user behavior, and in the back-end to generate appropriate behavior in turn. Dialog management models, for example turn-taking and argumentation, are part of the interaction models.

In the remainder of this paper, we discuss and analyze the current state of the art in HCI systems and the evaluation thereof. However, the challenges that we are facing when dealing with evaluation of emerging HCI applications are discussed in the context of Human Computing since we feel that proper models of interaction are essential for these applications.

3 HCI Systems

3.1 Traditional HCI Systems

Traditional HCI systems allow human users to input commands using keyboards, mice or touch screens (e.g. ATM machines, web browsers, online reservation systems). These input devices are reliable in the sense that they are unambiguous. Traditionally, systems are single-user, task-oriented and the place and manner in which the interaction takes place are largely determined by the projected task and expected users. This allows system designers to specify the syntax and style of the interaction. Since both input and output interfaces are physical, an explicit dialogue between the user and the computer

can be established. This dialogue is more even more facilitated when only the user can initiate the interaction.

If we look at the ATM, the user that interacts with the system clearly wants to perform a task: withdrawing money or viewing the account balance. The interaction devices are physical: the buttons and card reader for input into the system, the screen, money slot and ticket printer for output to the user. ATMs are intended to be single-user, and the user always initiates the interaction. The dialogue between user and system is explicit, and highly standardized.

3.2 Emerging HCI Systems

Emerging HCI systems and environments have a tendency to become *multi-modal* and *embedded* and thereby allowing people to interact with them in natural ways. In some cases, the design of computer interfaces is merging with the design of everyday appliances where they should facilitate tasks historically outside the normal range of human-computer interaction. Instead of making computer interfaces for people, people have started to make people interfaces for computers [5].

The nature of applications is changing. Looking beyond traditional productivity-oriented workplace technologies where performance is a key objective, HCI is increasingly considering applications for everyday life. HCI design now encompasses leisure, play, culture and art. Compared to traditional HCI systems, we can identify four main trends in HCI systems:

1. **New sensing possibilities.** New sensing technologies allow for the design of interfaces that go beyond the traditional keyboard and mouse. Automatic speech recognition is common in many telephone applications. The current state of video tracking allows not only for localization of human users, but also to detect their actions, identity and facial expressions [4]. This opens up possibilities to make interfaces more natural. Humans will be able to interact in ways that are intuitive. However, this comes at a cost of having to reconsider the syntax of the application. When using speech or gestures, the vocabulary is almost infinite. Moreover, many of the 'behaviors' that we recognize, must be interpreted in relation to the context. Context aware applications employ a broad range of sensors such as electronic tags, light sensing and physiological sensing. However, integration and the subsequent interpretation of these signals is hard, and context aware systems are likely to consider contexts differently than users do [6]. For example, when a user decides to watch a movie at home and closes the blinds to make the room darker, the system may automatically switch on the lights. Clearly, there user and system have a different view of the current situation.

   Related to the use of a multiplicity of sensors is the trend that sensors are moving to the background [7; 8]. This moves interfaces away from the object-oriented approach that is traditionally considered [9]. This trend has large implications for interaction design since it restricts the traditional dialog-oriented way of interaction, and effort must be paid to the design of implicit interactions [10].

2. **Shift in initiative.** Traditional HCI systems embrace the explicit way in which the dialog with the user is maintained. Moreover, the user is virtually always the one

who initiates the interaction. Consequently, traditional HCI systems are responsive in nature. Nowadays, pro-active systems are more common. Ju and Leifer [10] define an initiative dimension in their framework for classifying implicit interactions. They state that, when regarded more generally, there is direct manipulation at the one end, and autonomy at the other. They argue that for HCI, neither of these states are appropriate. Instead, the interaction is likely to be mixed-initiative. This implies that there must be a way to coordinate the interaction, which should be the focus of interaction design. For applications where multiple users can engage in the interaction, there's also a mixed-initiative among the different users.

3. **Diversifying physical interfaces.** The physical forms of interfaces are diversifying [11], as was foreseen by Mark Weiser [8]. One movement is to make interfaces bigger, such as immersive displays and interactive billboards. Another movement is to make interfaces smaller, such as wearable and embedded displays. This last movement is largely motivated by the popularity of mobile devices. The market for mobile phones is still growing, and so is the number of applications. With the increased connectivity and bandwidth, it is possible that people interact remotely with the same application. The trend of diversifying physical interfaces is most visible for general purpose desktop computers. These are increasingly often replaced by more purpose-designed and specialized appliances [11].

4. **Shift in application purpose.** There is a shift in application purpose for HCI systems. This shift is partly a consequence of new technology, and partly motivates the development of new technology. Whereas traditional systems are, in general, task-based, new applications are more focused on everyday life [11], thus on the user. User Experience (UX), although associated with a wide variety of meanings [12], can be seen as the countermovement of the dominant task and work related 'usability' paradigm.

   UX is a consequence of a user's internal state (e.g. predispositions, expectations, needs, motivation and mood). The literature on UX reveals three major perspectives [13]: human needs beyond the instrumental; affective and emotional aspects of interaction; and the nature of experience. Hassenzahl and Sandweg [14] argue that future HCI must be concerned about the pragmatic aspects of interactive products as well as about hedonic aspects, such as stimulation (personal growth, increase of knowledge and skills), identification (self-expression, interaction with relevant others) and evocation (self maintenance, memory). The task is no longer the goal, but rather the interaction itself (e.g. [15]).

   Typical UX applications are focused on leisure, play, culture and art. Consequently, this focus affects the interface. Factors as pleasure, aesthetics, expressiveness and creativity play an increasingly important role in the design of both interface and interaction. Video games are a clear example of UX applications.

   Another aspect is that interfaces are not only more centered on the user and the interaction, but also show a trend towards product integration. Domestic technology is becoming increasingly complex [16]. Our microwaves function also as stoves, we can listen to music, take pictures and exchange media with our mobile phones and our washing machines can also dry the laundry for us. Ubiquitous computing (UC), although radically different from traditional HCI on a number of criteria, is

one extreme example where functionality is integrated. Of course, this influences the choice of physical interface.

4 Stressing the Need for Evaluation: Three Examples of Emerging HCI Applications

In this section, we discuss three examples of emerging HCI systems. These serve to demonstrate the observed trends in HCI system development, and allow us to pinpoint the difficulties with traditional evaluation methods in Section 5.3.

4.1 Groupware Systems

One example of an area where a lot of money has been invested into the development of a product because of its expected scenario gains is the area of group support systems (GSS) or groupware. De Vreede *et al.* [17] conclude from extensive research that 15 years after the introduction of the first group support system, these systems indeed provide added value to meetings. They are said to provide savings, and increase efficiency. It was a rather complex and non-straightforward process to come to this conclusion.

One of the reasons that it took so long was the fact that people were facing difficulties when using the system, as they were not familiar with the changes in work practice that were introduced by them [18]. People were forced to use novel tools during meetings and had to abandon their common meeting practice. As a consequence, also the benefits proved hard to measure as people objected to the use of these tools.

GSS are clear examples of systems that establish a *shift in application purpose*. Although Grudin [19] already noted that adequate understanding of the political and social factors at work were to be considered in the design and implementation phases in order to avoid an initial reject from the public, the task of supporting the meeting process (e.g. facilitate brainstorming) was considered more important than how these systems were used in practice. It was therefore not strange that people found it difficult to understand what the system was supposed to do for them and their group [20]. Design for intuitive interaction with the user as focal point would have facilitated its adoption, without any doubt.

4.2 Smart Homes

Smart home systems are typical examples of ubiquitous systems, characterized by their pervasive nature. Users are observed in their homes using a large number of sensors, ranging from cameras and microphones to pressure and heat sensors. See Figure 2(a) for an example of a smart home setting. From a user point of view, ubiquitous systems do not necessarily have a task. They can be anywhere between responsive and pro-active. An example that lies somewhere in between responsive and pro-active is for instance the smart home described in Intille *et al.* [6] where the system *suggests* users which clothes to wear given the outside temperature, or suggests measures to save energy. From a system point of view, smart homes have the task to maintain the homeostasis of the environment, and to support the users that are living in it. One example is a smart

(a) (b)

Fig. 2. (a) Philips' vision of smart homes: the environment is adjusted to make patients feel more comfortable in hospitals. (b) User interacting with the Virtual Dancer.

home that supports elderly people in order to allow them to live (semi-)independently. These homes not only take care of lighting and heating issues, but also facilitate communication in case of emergencies.

When the environment itself becomes the interface, people go about their daily lives and perform their tasks while the computing technologies are there to support them transparently [8]. People start to implicitly interact with computers and technology that have moved to the background. Despite being written over 10 years ago, many aspects of Mark Weiser's vision of ubiquitous computing appear as futuristic today as they did in 1991 [21].

As Davies and Gellersens [22] mention there are many aspects that need to be resolved before ubiquitous interfaces really will break through. They mention, amongst others, the need for fusion models and context awareness. Due to the lack of an explicit interface, users are required to communicate naturally with the system. This requires fusion of multiple communication channels. The system must be aware of the context, and interpret the user's actions in this context. On the other hand, the user must be familiar with the system's abilities, and system's state.

The complexity and black box characteristics of smart homes make them even more difficult to evaluate. They do not only introduce a *shift in application purpose*, but also employ *new sensing possibilities*. There is a radical *change in physical interface* since the smart home has become the interface itself. Some smart homes are pro-active, which presents a clear *shift in initiative*.

4.3 Virtual Dancer

Fun and entertainment are becoming increasingly important in almost all uses of information technology [23]. *Ambient entertainment* is the field of research that deals with applications that are centered around this theme. One example of an ambient

entertainment application is the Virtual Dancer, as described in Reidsma *et al.* [15]. It is an interactive installation where users can dance together with a virtual character. The virtual character reacts to the observed movements of the user, and tries to influence the movements of the user in turn. The movements are observed using a dance mat and a camera. The camera recognizes global movement features that are mapped onto a database of prerecorded movements for the Virtual Dancer.

The camera and dance mat provide *new sensing possibilities*. Also, during the dance, there is a constant *shift in initiative*. The goal of the application is to entertain the user, without the provision of an explicit task. Instead, the interaction itself is the goal of the application, a clear *shift in application purpose*.

In this so-called taskless interaction, not the task but the interaction itself and the user experience need to be evaluated. Attempts so far to evaluate the interaction have been limited to analyzing video recordings of the user in order to determine engagement in the interaction. This does not allow for reliable assessment of aspects that improve the user's experience during the interaction, let alone which parts of the system should be improved. One important aspect is that the responses of the user to certain actions of the systems have to be measured. This requires the knowledge of system states, i.e. the context. While this information proves valuable in the assessment of the participation level of the user, it does not provide much information about the actual user experience. Instead, this information could be collected using questionnaires or by employing biosensors that measure heart rate and the respiratory level.

5 Evaluation

Evaluation is broad concept. In the domain of HCI, Preece *et al.* [24], page 602 defined this concept in 1994 as follows:

> Evaluation is concerned with gathering data about the usability of a design or product by a specific group of users for a particular activity within a specified group of uses or work context.

In 2007 they have expanded this definition [25], page 584:

> It [evaluation] focuses on both the usability of the system, e.g. how easy it is to learn and to use, and on the users' experience when interacting with the system, e.g. how satisfying, enjoyable, or motivating the interaction is.

The use of evaluation methods for the assessment of the suitability of HCI systems has become a standard tool in the design process. Many HCI systems are designed iteratively, where in each cycle design issues of the previous one are addressed. These issues are identified in an evaluation step. We discuss the design criteria of HCI systems first in Section 5.1. We then focus on current evaluation practice in the HCI field in Section 5.2. Section 5.3 discusses issues that appear when dealing with evaluation for emerging HCI applications.

5.1 Design Criteria in HCI

Much has been written about the design of HCI systems (e.g. [26; 27]). Designed well, interactive systems can allow us to reap the benefits of computation and communication away from the desktop, assisting us when we are physically, socially or cognitively engaged, or when we ourselves do not know what should happen next. Designed poorly, these same devices can wreck havoc on our productivity and performance, creating irritation and frustration in their wake [10]. Good practice is to explicitly formulate design choices.

Norman [28] identifies a number of principles for good interaction design. Often used are the principles visibility, feedback, constraints, consistency, recovery, and affordance. If things are *visible*, people can see what functions are available and what the system is currently doing. Then they are more likely to know what to do next as a consequence of the psychological principle that it is easier to recognize things than to recall them. The *feedback* principle is related to the visibility principle and refers to sending back information from the system to the users so that they know what effect their actions have had. Timely feedback provides the necessary visibility for user interaction and will enhance the feeling of control. *Constraints* should prevent people from making errors through properly constraining allowable actions. For instance by deactivating certain menu options people are restricted to choose only permissible actions. *Consistency* is a design principle that emphasizes the importance of uniformity in the placement, appearance and behavior of screen elements and operations to make systems easier to learn and use. Recovery refers to the principle that a system should enable users to recover from actions, particularly errors and mistakes, quickly and effectively. Finally, the principle of *affordance* refers to the fact that things should be designed in a way that it is clear what they are for and how to use them. For instance buttons afford pressing and should look like buttons to invite people to press them.

Traditionally, HCI systems are designed for a certain task, in a given context, and with a certain user profile in mind. Key point is that the HCI system must be useful, usually referred to as usability. There are many different approaches to making a product usable and there is no generally accepted definition. Nielsen [29] identifies five components of usability that are commonly used: efficiency, learnability, memorability, errors, and satisfaction. Three of these criteria can be used as quantitative indicators to asses the usability of a system by measuring respectively time to complete a task, time to learn a task and the number of errors made when carrying out a specific task. Other important usability goals are effectiveness and utility referring to how good a system is at doing what it is supposed to do and to what extent the system provides the right kind of functionality [25].

In addition, usability can be regarded from three distinct viewpoints [30; 31]: product-oriented, user-oriented and user performance-oriented. The product-oriented view can be measured in terms of ergonomic attributes of the product. The user-oriented view in terms of mental effort and attitude of the user and the user performance-view by examining how the user interacts with the product with emphasis on either the ease of use or the acceptability of the product in the real world.

The above views are complemented by the contextual view, which tells us that usability of a product is a function of a particular user class of users being studied, the application at hand and the environment in which they work.

Besides usability, in the interaction between the human and the computer also the user interface and user experience come into play. To stress the necessity of shifting the focus from what computers can do to what users can do, Shneiderman introduced the term "the new computing" [32]. The broadening of the user community to include almost everyone, even technology resisters, puts a challenge on system designers who realize now that understanding the user is important. This forms the basis of User-Centered Design (UCD). UCD is a multidisciplinary design approach based on the active involvement of users to improve the understanding of user and task requirements, and the iteration of design and evaluation [33]. It has been mentioned that this approach is the key to product usefulness and usability and overcomes the limitations of traditional system-centered design [32].

One view of UCD is to design HCI as close as possible to natural human-human interaction [34]. The rationale is that users do not have to learn new communication protocols, which leads to increased interaction robustness. This aids the user experience and provides guidelines for designing the user interface. A drawback is that one should be familiar with the application to know what to expect from it. Shneiderman [27] argues that most designs for natural interaction do not provide users with available task actions and objects. In terms of the design principles treated before they violate visibility. For knowledgeable and frequent users who are aware of available functions this will not be a problem but for them a precise, concise command language is usually preferable. Hence Shneiderman claims that natural interaction will only be effective for intermittent users who are knowledgeable about specific tasks and interface concepts but have difficulties remembering syntactic details.

5.2 Current Evaluation Practice in HCI

As stated before, evaluation is nowadays common practice in the field of HCI. The use of evaluation methods is motivated by the reported increased return on investments.

In general, we can identify two broad classes of evaluation methods: expert-based evaluation (e.g. cognitive walkthrough, heuristic evaluation, model based evaluation) and user-based evaluation (e.g. experimental evaluation, user observation, use of questionnaires, monitoring physiological responses). The bulk of early HCI designers and evaluators were cognitive psychologists. Cognitive models like GOMS [35] were very influential, as were laboratory experiments. Nielsen [29] took a more pragmatic approach, stating that full-scale evaluation of usability is too complicated in many cases, so that 'discount' methods are useful instead. His work has been very influential, partly due to the ease of application, partly due to the relative low cost. His vision has lead to an enormous number of different methods in regular use for the evaluation of usability.

Since its early days, HCI research focussed almost exclusively on the achievement of behavioral goals in work settings. The task that had to be performed by the user was the pivotal point of user centered analysis and evaluation. Rengger [36] defined four classes of performance measures:

1. Goal achievement (accuracy and effectiveness)
2. Work rate (productivity and efficiency)
3. Operability (function usage)
4. Knowledge acquisition (learning rate)

Though satisfaction has been one of the components of usability since the early days, in the last few years there is an increased focus on user experience. The expansion of the definition of evaluation in the beginning of Section 5 is not typical for the book of Sharp, Rogers and Preece [25]. Similar changes can be found in most literature on HCI. This is one of the answers of the HCI community to the shift in application purpose of systems mentioned in Section 3.2. But as we discussed before, emerging HCI systems require other measures, and other evaluation practice. In the next section, we identify challenges for evaluation of emerging HCI systems, and use the examples in Section 4 as an illustration.

5.3 Challenges for Evaluation of Emerging HCI Systems

The characteristics of emerging HCI systems imply that traditional approaches to usability engineering and evaluation are likely to prove inappropriate to the needs of its users. As a result of the trends that we discussed in Section 3.2, problems emerge in the design and evaluation of HCI systems. We start by discussing the front-end of Human Computing, using the examples of Section 4.

Human Sensing. The use of keyboards, buttons and mice for interaction with HCI systems is found to be inconvenient since these devices do not support the natural ways in which humans interact. Although debated, the use of natural communication is often considered more intuitive, and therefore expected to be more efficient from a user's point of view. Voice, gestures, gaze and facial expressions are all natural human ways of expression. In natural contexts, humans will use all these channels, one to enhance and complement an other. To make truly natural interfaces, this implies that all these channels should be taken into account. This, however, is difficult for at least three reasons:

1. The recognition is error-prone
2. The lexicon of expression is much larger than with 'artificial input'
3. Integration of multiple channels often leads to ambiguities

Error-prone recognition. When using natural channels, the data obtained from sensors (microphone, camera) needs to be analyzed. From the streams of data, we need to recognize the communicative acts (e.g. words, gestures, facial expressions). Although much research is currently devoted to making automatic recognition more accurate, these systems will never be error-free. Another aspect is that automatic recognition is probably less fine-grained than what human observers are able to perceive [37]. Subtleties might easily go unnoticed.

Reduction of errors is probably the most convenient way of improving the usability. However, as recognition will never be error-free, repair mechanisms need to be present.

Feedback and insight in the system state are useful because they give the user insight in how the input is recognized and interpreted. Still, there are many challenges in how to present the feedback or system state [38]. One approach in speech recognition tasks is to feed the recognition back to the user. The can be done, for example, in applications where tickets can be ordered via telephone. After the user has specified the date, the system could ask "How many tickets do you want to order for this Tuesday?" Although this kind of mechanism can improve the recognition performance, attention must be paid to how users can correct the recognition.

Assessment of the input reliability is an important aspect of usability evaluation. One way to do this is by applying standard benchmark sets. Well-known benchmark sets are the NIST RT sets [39] for automatic speech recognition or FRVT and FRGC for face recognition [40]. These sets are specific for a given context and task. Since they contain ground truth and the error metrics are known, they allow for good comparison of recognition algorithms. However, they still evaluate only the reliability of the input. In addition to this, the system must be evaluated together with the (unreliable) input.

Large lexicon. In natural human-human interaction, humans use a large lexicon of speech, and eye, head and body movements, both conscious and unconscious. When allowing humans to communicate with HCI systems in a natural way, the input devices should be able to recognize the whole range of signals. This poses severe requirements on the recognition.

Two factors are important when evaluating the lexicon. First, the lexicon should be sufficiently large to allow for the recognition of all foreseen (and unforeseen) actions. For a system such as the Virtual Dancer (see Section 4.3), this implies that the whole range of dance movements that a user can make, should be included in the lexicon. Alternatively, only a subset of all communication signals can be considered. However, it should be acknowledged to the user whether the signals are recognized and interpreted.

Second, the choice of the lexicon should be intuitive. In many cases, an *ad hoc* lexicon is chosen, often to maximize the recognition. Ideally, the lexicon should contain signals that users naturally make when interacting with the HCI system. Note that, although this interaction is natural, the lack of a clear interface might prove that it is also not intuitive [41]. A preliminary investigation should be conducted to see what these movements and sounds are, for example by conducting Wizard of Oz experiments. An example of such an investigation is described in [42].

When dealing with attentive or pro-active systems, not only the communicative actions are of importance. These systems require awareness of things as user state and intentions, which generally can be deducted from behavior that is non-communicative.

Integration of channels. Human behavior is multi-modal in nature. For example, humans use gestures and facial expressions while speaking. Understanding of this behavior does not only require recognition of the input of individual channels, but rather the recognition of the input as a whole. Despite considerable research effort in the field of multi-modal fusion (see e.g. [43]), our knowledge about how humans combine different channels is still limited. When dealing with multi-user systems, the problem is even harder since also the group behavior needs to be understood. For remote participation in the interaction, it might be very difficult since the interaction mechanisms for

human-human interaction are probably not applicable. Furthermore, due to the disappearing interfaces, the lack of explicit turn-taking will cause users to employ many alternate sequences of input, and requires HCI systems to be more flexible in handling these in turn [9].

Similar to the performance evaluation of single communication channels, the recognition of the fused channel information need to be assessed. Integration of multiple channels can lead to reduction of signal ambiguity, provided that the context is known. Therefore, accurate assessment of the context is needed.

Context Awareness. It is often mentioned that human behavior is to be interpreted in a given context. For example, a smile in a conversation can be a sign of appreciation, whereas, during negotiation, it can show disagreement. So for reliable interpretation of the human behavior, it is important to be aware of the context of the situation. To date, there is no consensus of what context is precisely, and how we should specify this [44]. Without a good representation for context, developers are left to develop *ad hoc* and limited schemes for storing and manipulating this key information [37]. This is acceptable for small domains, but is inappropriate for larger and more complex applications.

Usually, the context is specified as the identity and location of the users, and the characteristics and timing of the action performed. Ideally, even the intentions of the user should also be taken into account. This is particularly difficult since these cannot be measured. These components of context are referred to as the 5 Ws [37; 4]: who, what, where, when, why. These basic components are limited, and one might include the identity and locations of all objects of interest, as well as the current goal of the user. Also, the history of all environment changes and user actions are considered important for reasoning about the context.

It difficult to assess the right values for all these properties, and context aware systems are likely to consider contexts differently than users do. Intille *et al.* [6] observe that, for smart homes (see Section 4.2), the user naturally considers contexts that the system has not, and propose to use suggestive systems, rather than pro-active ones.

Reference Tasks. Whittaker *et al.* [45] observed that many developed HCI systems can be considered radical inventions. They do not build further on established knowledge about user activities, tasks and techniques but rather push the technology envelope and invent new paradigms. Although we lack basic understanding of current users, tasks and technologies, the field of HCI is encouraged to try out even more radical solutions, without pausing to do the analysis and investigation required to gain systematic understanding. The absence of shared task or goal information makes it difficult to focus on research problems, to compare research results and to determine when a new solution is better, rather than different. This prevents proper consolidation of knowledge.

When the users are not familiar with the task or goal the application supports, users are likely to use the system in a different way. This makes evaluation of the fitness of the system difficult. For example, interfaces that support creative thinking are designed for a specific task that is new to the users. Without proper familiarization, these interfaces are less effective (see for example the Groupware example in Section 4.1).

The lack of reference tasks can be seen as a challenge for the development of proper interaction models. Now we move to the back-end of Human Computing.

Performance Metrics. In contrast to Rengger [36], as discussed in Section 5.2, emerging HCI applications often do not have well-defined tasks, which asks for novel measures. There are many factors in HCI that have a substantial impact on the success of applications but are not easily quantified. Amongst them are user experience [16], fun [46], ethical issues [47], social relationships [48] and aesthetical issues [1]. For example, for the Virtual Dancer (see Section 4.3), it remains a challenge to define proper measures to evaluate the success of the interaction. These critical parameters are also required in order to compare similar applications [49]. When the application supports multiple users, these measures might be shared among the users.

Learnability. Given the increasing complexity of HCI systems, it is to be expected that the time needed to learn to work with a system grows along. Currently, evaluation of these systems focuses on 'snap shots', but fail to focus on the learning [50]. Longitudinal studies that assess how the use of a system develops from the first encounter are needed to gain insight in what kind of barriers users encounter when using the system, and how they solve these.

Context of Authentic Use. HCI systems should be evaluated in a context as close as possible to the context of authentic use [37]. The context is often difficult to realize, especially for multi-user applications. Evaluating HCI systems in laboratory settings is likely to cause unnatural behavior of the users. This makes proper evaluation of the system difficult, if not impossible.

Another drawback of using laboratory testing is that parameters can be controlled (background noise, lightning conditions) that cannot be controlled in the context of authentic use. As a consequence, there is a difference in how these systems perform in reality.

As an example, the live-in laboratory PlaceLab [51] has been built to ensure that assumptions about behavior in the lab correspond to behavior in more realistic (and complex) situations in real smart homes.

6 Conclusion

New HCI systems are emerging that differ from traditional single-user, task-based, physical-interface HCI systems. We identify four trends: new sensing possibilities, a shift in initiative, diversifying physical interfaces, and a shift in application purpose. Traditional evaluation practice does not suffice for these new trends.

The use of more natural interaction forms poses problems when the input is ambiguous, the communication lexicon is potentially large, and when interpreting signals from multiple communication channels, ambiguities might arise. Identifying the context of use is important because interpretation of input is often dependent on the context. For complex systems, sensing the context is increasingly difficult. Evaluation of context aware systems is consequently difficult.

There is no consensus about appropriate performance metrics for emerging HCI systems. Task-specific measures are useless for evaluation of task-less systems. Related to this is the lack of common reference tasks. The 'radical invention' practice in the field

of HCI prevents proper consolidation of knowledge about application tasks and goals, and user activities. Therefore, it is difficult to compare HCI systems.

As HCI systems are becoming more complex, the learning process of users is more and more important. This is currently a neglected part of evaluation. The introduction of longitudinal evaluation studies is needed to gain insight in the learning mechanisms. A final practical issue is the lack of authentic usage contexts. Many systems are only evaluated in a laboratory setting, instead in their projected context.

We summarized trends in HCI systems and pointed out where problems appear. We discussed three examples of complex HCI systems, and argued the need for appropriate evaluation. With this paper, we aimed at achieving increased awareness that evaluation too has to evolve to support the emerging trends in HCI systems.

Acknowledgments

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access, publication AMIDA-2), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein. The authors wish to thank Anton Nijholt for his valuable comments.

References

[1] Alben, L.: Quality of experience: defining criteria for effective interaction design. Interactions **3**(3) (1996) 11–15

[2] Gaver, B., Martin, H.: Alternatives: exploring information appliances through conceptual design proposals. In: Proceedings of the conference on Human factors in computing systems (CHI'00), The Hague, The Netherlands (2000) 209–216

[3] Jaimes, A., Sebe, N., Gatica-Perez, D.: Human-centered computing: A multimedia perspective. In: Proceedings of the ACM international conference on Multimedia, Santa Barbara, CA (2006) 855–864

[4] Pantic, M., Pentland, A., Nijholt, A., Huang, T.S.: Machine Understanding of Human Behavior: A Survey. In: Artificial Intelligence for Human Computing. Volume 4451 of Lecture Notes in Artificial Intelligence Springer-Verlag (2007) 47–71

[5] Coen, M.H.: Design principles for intelligent environments. In: Proceedings of the National Conference on Artificial Intelligence (AAAI'98), Madison, WI (1998) 547–554

[6] Intille, S.S., Tapia, E.M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S., Bao, L., Larson, K.: Tools for studying behavior and technology in natural settings. In: Proceedings of the International Conference on Ubiquitous Computing (Ubicomp'03). Volume 3869 of Lecture Notes in Computer Science., Seattle, WA (2003) 157–174

[7] Streitz, N., Nixon, P.: Introduction: The disappearing computer. Communications of the ACM **48**(3) (2005) 32–35

[8] Weiser, M.: The computer of the 21st century. Scientific American **265**(3) (1991) 66–75

[9] Nielsen, J.: Noncommand user interfaces. Communications of the ACM **36**(4) (1993) 83–89

[10] Ju, W., Leifer, L.: The design of implicit interactions. Design Issues, Special Issue on Design Research in Interaction Design (to appear)

[11] Benford, S., Schndelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T., Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S., Schmidt, A., Gellersen, H.W., Steed, A.: Expected, sensed, and desired: A framework for designing sensing-based interaction. ACM Transactions on Computer-Human Interaction **12**(1) (2005) 3–30

[12] Forlizzi, J., Battarbee, K.: Understanding experience in interactive systems. In: Proceedings of the conference on Designing Interactive Systems (DIS'04), Cambridge, MA (2004) 261–268

[13] Hassenzahl, M., Tractinsky, N.: User experience: a research agenda. Behaviour & Information Technology **25**(2) (2006) 91–97

[14] Hassenzahl, M., Sandweg, N.: From mental effort to perceived usability: transforming experiences into summary assessments. In: Extended abstracts on Human factors in computing systems (CHI'04), Vienna, Austria (2004) 1283–1286

[15] Reidsma, D., van Welbergen, H., Poppe, R., Bos, P., Nijholt, A.: Towards bi-directional dancing interaction. In: International Conference on Entertainment Computing (ICEC'06). Volume 4161 of Lecture Notes in Computer Science. (2006) 1–12

[16] Thomas, P., Macredie, R.D.: Introduction to the new usability. ACM Transactions on Computer-Human Interaction **9**(2) (2002) 69–73

[17] de Vreede, G.J., Vogel, D.R., Kolfschoten, G.L., Wien, J.: Fifteen years of GSS in the field: A comparison across time and national boundaries. In: Proceedings of the Hawaii International Conference on System Sciences (HICSS'03), Big Island, HA (2003) 9

[18] Nunamaker Jr., J.F., Briggs, R.O., Mittleman, D.D.: Electronic meeting systems: Ten years of lessons learned. In: Groupware: Technology and Applications. Prentice Hall, Englewood Cliffs, NJ (1995)

[19] Grudin, J.: Groupware and social dynamics: Eight challenges for developers. Communications of the ACM **37**(1) (1994) 93–105

[20] Briggs, R.O., de Vreede, G.J., Nunamaker Jr., J.F.: Collaboration engineering with thinklets to pursue sustained success with group support systems. Journal of Management Information Systems **19**(4) (2003) 31–64

[21] Schmidt, A., Kranz, M., Holleis, P.: Interacting with the ubiquitous computer: towards embedding interaction. In: Proceedings of the joint conference on Smart objects and ambient intelligence (sOc-EUSAI'05), Grenoble, France (2005) 147–152

[22] Davies, N., Gellersens, H.W.: Beyond prototypes: Challenges in deploying ubiquitous systems. IEEE Pervasive Computing **2**(1) (2002) 26–35

[23] Wiberg, C.: Usability and fun: An overview of relevant research in the hci community. In: Proceedings of the CHI Workshop on Innovative Approaches to Evaluating Affective Interfaces, Portland, OR (2005)

[24] Preece, J., Rogers, Y., Sharp, H., Benyon, D.: Human-Computer Interaction. Addison-Wesley Longman Ltd. (1994)

[25] Sharp, H., Rogers, Y., Preece, J.: Interaction Design: Beyond Human Computer Interaction. 2nd edn. John Wiley and Sons (2007)

[26] Dix, A., Finlay, J., Abowd, G., Beale, R.: Human Computer Interaction, third edition. Prentice Hall (2004)

[27] Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. 4th edn. Addison Wesley (2005)

[28] Norman, D.A.: The Design of Everyday Things. MIT Press (1998)

[29] Nielsen, J.: Usability Engineering. Academic Press, Boston, MA (1993)

[30] Bevan, N., Kirakowski, J., Maissel, J.: What is usability? In: Proceedings of the international Conference on HCI, Stuttgart, Germany (1991) 651–655

[31] Rauterberg, M.: Quantitative measures for evaluating human-computer interfaces. In: Proceedings of the International Conference on Human-Computer Interaction, Orlando, Florida (1993) 612–617
[32] Shneiderman, B.: Leonardo's Laptop: Human Needs and the New Computing Technologies. MIT Press (2002)
[33] Mao, J.Y., Vredenburg, K., Smith, P.W., Carey, T.: The state of user-centered design practice. Communications of the ACM **48**(3) (2005) 105–109
[34] Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S.L., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T., Stanney, K.M., Su, H., Wang, Q.Y.: Guidelines for multimodal user interface design. Communications of the ACM **47**(1) (2004) 57–59
[35] Card, S.K., Newell, A., Moran, T.P.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Mahwah, NJ (1983)
[36] Rengger, R.E.: Indicators of Usability based on performance. In: Human Aspects in Computing: Design and Use of Interactive Systems with Terminals. Elsevier, Amsterdam, The Netherlands (1991) 656–660
[37] Abowd, G.D., Mynatt, E.D.: Charting past, present, and future research in ubiquitous computing. ACM Transactions on Computer-Human Interaction **7**(1) (2000) 29–58
[38] Bellotti, V., Back, M., Edwards, K., Grinter, R.E., Jr., A.H., Lopes, C.V.: Making sense of sensing systems: Five questions for designers and researchers. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI'02), Minneapolis, MN (2002) 415–422
[39] Fiscus, J.G., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C.: The rich transcription 2005 spring meeting recognition evaluation. In: Revised Selected Papers of the Machine Learning for Multimodal Interaction Workshop 2005 (MLMI'05). Volume 3869 of Lecture Notes in Computer Science., Edinburgh, United Kingdom (2006) 369–389
[40] Phillips, J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Worek, W.: Preliminary face recognition grand challenge results. In: Proceedings of the Conference on Automatic Face and Gesture Recognition 2006 (FGR'06), Southampton, United Kingdom (2006) 15–24
[41] Nijholt, A., Rist, T., Tuijnenbreijer, K.: Lost in ambient intelligence? In: Extended abstracts on Human factors in computing systems (CHI'04), Vienna, Austria (2004) 1725–1726
[42] Höysniemi, J., Hämäläinen, P., Turkki, L., Rouvi, T.: Children's intuitive gestures in vision-based action games. Communications of the ACM **48**(1) (2005) 44–50
[43] Oviatt, S.L.: 14: Multimodal interfaces. In: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Associates (2003) 286–304
[44] Van Bunningen, A.H., Feng, L., Apers, P.M.: Context for Ubiquitous Data Management. In: International Workshop on Ubiquitous Data Management (UDM'05), Tokyo, Japan (2005) 17–24
[45] Whittaker, S., Terveen, L., Nardi, B.A.: Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI. Human Computer Interaction **15**(2-3) (2000) 75–106
[46] Blythe, M.A., Overbeeke, K.J., Monk, A.F., Wright, P.C.: Funology: From Usability to Enjoyment. Volume 3 of Human-Computer Interaction Series. Kluwer Academic Publishers (2003)
[47] Nardi, B.A., Kuchinsky, A., Whittaker, S., Leichner, R., Schwarz, H.: Video-as-data: Technical and social aspects of a collaborative multimedia application. Computer Supported Cooperative Work **4**(1) (1995) 73–100
[48] Grudin, J.: Why CSCW applications fail: problems in the design and the evaluation of organizational interfaces. In: Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'88), New York, USA (1988) 85–93

[49] Newman, W.M.: Better or just different? On the benefits of designing interactive systems in terms of critical parameters. In: Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, Amsterdam, The Netherlands (1997) 239–245

[50] Petersen, M.G., Madsen, K.H., Kjær, A.: The usability of everyday, technology-emerging and fading opportunities. ACM Transactions on Computer-Human Interaction **9**(2) (2002) 74–105

[51] Intille, S.S.: The goal: smart people, not smart homes. In: Proceedings of the International Conference on Smart Homes and Health Telematics, Belfast, United Kingdom (2006) 3–6

Gaze-X: Adaptive, Affective, Multimodal Interface for Single-User Office Scenarios

Ludo Maat[1] and Maja Pantic[2,3]

[1] EEMCS, Delft University of Technology, Delft, The Netherlands
[2] Computing Dept., Imperial Collge London, London, UK
[3] EEMCS, University of Twente, Enschede, The Netherlands
ludomaat@zonnet.nl, m.pantic@imperial.ac.uk

Abstract. This paper describes an intelligent system that we developed to support affective multimodal human-computer interaction (AMM-HCI) where the user's actions and emotions are modeled and then used to adapt the interaction and support the user in his or her activity. The proposed system, which we named Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how). It integrates a number of natural human communicative modalities including speech, eye gaze direction, face and facial expression, and a number of standard HCI modalities like keystrokes, mouse movements, and active software identification, which, in turn, are fed into processes that provide decision making and adapt the HCI to support the user in his or her activity according to his or her preferences. A usability study conducted in an office scenario with a number of users indicates that Gaze-X is perceived as effective, easy to use, useful, and affectively qualitative.

Keywords: Human Sensing, Context Sensing, Multimodal Interfaces, Affective Computing, Anthropocentric Interface Design.

1 Why Multimodal, Context-Sensitive User Interface Design?

We have entered an era of pervasive computing. Computers and the Internet have become so embedded in the daily fabric of our lives that we can no longer live without them [15]. We use them to work, study, communicate, shop, and entertain ourselves. With the ever-increasing diffusion of computers into society, human-computer interaction (HCI) is becoming increasingly essential to our daily lives.

Predicting the future of HCI is a difficult task, but one important source of help is the accumulated information about the preferences and limitations of humans interacting with computers. Principles can be drawn upon, which may explain why some interfaces survive and others become extinct. Rigid designs that assume that users will be explicit and fully attentive while interacting with the computer, that do not protect against errors, provide help at all times except at the right moment, and all in all make users frustrated, are likely to become quickly extinct due to their poor usability [28]. On the other hand, designs that include adequate attention to individual differences among users, support (natural) multimodal and context-sensitive

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 251–271, 2007.

interaction, expend on designs for reliability and safety, provide access to the elderly and handicapped, and properly adapt to the user level of knowledge, skills, attention, preferences, moods, and intentions, are the kind of HCI designs that are likely to become the trend in computing technology [4], [41], [48], [39], [38], [51]. Although this list may not be complete, it points out important issues that are rather insufficiently addressed by the current initiatives [43].

1.1 The Evolution of Human-Computer Interfaces

Around 1980, at the dawn of the personal computer age, many chaotic and rigid user interfaces were produced that turned the users into frustrated victims of machines they could not control. Typical examples of useless interfaces at that time could display a five-minute video without a stop button and generate choice sequences that could not be reversed or cancelled. As high-resolution displays and fast chips emerged, video and audio processing as well as animations flourished (particularly for video games), giving rise to a new generation of user interfaces, in which direct manipulation became the dominant form of interacting and WYSIWYG (what you see is what you get) became a guiding principle. The aim was: (i) to make operations visible, incremental, rapidly manageable by means of a keyboard, and reversible, as well as (ii) to prevent user errors by effective designs. During the late 80s and early 90s, direct-manipulation interfaces were enhanced with embedded menus in text and graphics, mice, and various joysticks as the devices of choice.

As remarked by Preece and Schneiderman [41] and Pentland [37], the mid 90s can be viewed as the dawn of pervasive computing that shed a new light on the future of computing and gave rise to novel requirements that useful user interfaces should fulfill. The growing availability of World Wide Web access with embedded menus providing links across the world led to an unusually rapid growth of Web servers and applications. The necessity of delivering new products in an ever-decreasing time frame affected, consequently, the quality of the issued products and interfaces. This Internet hype also blurred the essence of some paradigms, such as software agents since an increasing number of vaguely related applications needed legitimacy and sought it under the umbrella of the “agents”. Though it was unfortunate at one hand and accompanied by numerous shoddy Web-oriented applications, on the other hand this Internet hype initiated rapidly accelerating progress in facilitating accessibility, speed, and reduction of error and failure rates. Moreover, it changed our view on computing and commerce [50]. Above all, it forecasted the type of working environments and information-communication spaces we are about to use in our everyday activities. It clearly indicated that in the future, with the aid of computers, we will carry out our daily tasks, communicate, and entertain ourselves in cyberspaces across distance, cultures and time. Of course, the specifics of such cyberspaces including virtual cyber worlds and of the related interfaces, which should facilitate easy and natural communication within those spaces and with the variety of embedded computing devices, are far from settled. In this LNAI Special Volume on AI for Human Computing many relevant issues are discussed and debated but all agree that before this new generation of pervasive computing can be widely deployed, the users should experience it as being human-centered and universally usable.

The designers of older technologies such as postal services, telephones, and television, have reached the ultimate goal of having products and services that are universally usable, but developers of computing technology cannot claim the same. Schneiderman [48], [49], reports an average of 5.1 hours per week wasted by the users while trying to use computers. Consequently, despite visible progress in accessibility, increase of speed, and reduction of the error and failure rates, the primary experience of many computer users is dissatisfaction or even frustration. Common problems include incompatibility (of file formats, applications' versions, screen sizes, etc.) and low speed (e.g., due to varying network bandwidths and processor speeds). Although these issues are not of the least importance, the crucial problem, which is primarily responsible for the users' anxiety and dissatisfaction, is incomprehensibility of many currently available software packages and Internet services [47]. As remarked by several researchers (e.g., [48], [49], [47]), a large majority of HCI systems available today assume users' proficiency in computing. However, for a casual user, these systems are often cumbersome, lacking the adaptability necessary to accommodate users with various levels of computing skill and experience. Furthermore, virtually all "classic" HCI systems tend to confine the user to a less natural, single-modal means of interaction (e.g., a mouse movement, pressing of a key, speech input, or hand motion). For example, to manipulate a virtual object with a typical HCI system, the user is usually required to select the object by employing mouse motion, then point with the mouse at a control panel to change the object's color. On the other hand, in a more natural setup, the user would point at the object with his finger and say: "Make it red". Integration of more than one natural modality into an interface would potentially overcome the current limitations of HCI systems: it would ease the need for specialized training and ease the information- and command-flow bottleneck between the user and the computer. Besides, recent data shows that a multimodal HCI can be an effective means for reducing uncertainty of single-modally sensed data (such as speech or hand motion), thereby improving robustness [43]. Although the incorporation of all features of human-human interaction (i.e,. an intricate interplay of thoughts, language, and non-verbal communicative displays) into human-computer interaction may be very complex and difficult to achieve, equipping HCI systems with a multimodal setup so that they can approach naturalness, flexibility and robustness of human-human communication will give them the potential to:

- transcend the traditional, cumbersome and rigid mouse/keyboard interaction, and
- yield a more effective and efficient information- and command-flow between the user and the computer system and, by that,
- approach universal usability.

Another challenge in fashioning universally usable HCI systems is to make them context sensitive. The key idea is to account for individual differences of the users and for the overall situation in which the user acts. For computing technology applications, context can be defined as any information that can be used to explain the situation that is relevant to the interaction between users and the application [10]. For

a single-user scenario, the six questions that summarize the key aspects of the computer's context with respect to its human user are as follows:

- *Who?* (Who the user is?)
- *Where?* (Where the user is? For single-user desktop-computer scenarios this context question is superfluous as it is known where the user is – either in front of the computer or not.)
- *What?* (What is the current task of the user?)
- *How?* (How the information is passed on? Which interactive signals / actions have been used?)
- *When?* (What is the timing of displayed interactive signals with respect to changes in the computer environment?)
- *Why?* (What may be the user's reasons to display the observed cues? Except of the user's current task, the issues to be considered include whether the user is alone and what is his or her affective state.)

Based upon the user's identity and the knowledge about his or her environment and current task, it would be possible to retrieve information about the importance of that task in the given environment, the user's skills in performing that task, and the user's overall preferences. Together with the sensed user's affective state, this information could be employed to define the following.

What form should the instruction manual have? For example, for a novel user a lucid tutorial could be provided, to an average user constructive help files could be offered, for an expert user compact notes could prove to be sufficient. Yet the kind of help files provided to the user should not be rigid; it should be determined based upon the overall user's preferences and his/her current affective state. Namely, users might prefer compact help files even if they are novices and especially if they are in a hurry.

When should the user be interrupted? This is of particular importance for the systems that are designed such that they rely on the user's feedback at all times or offer assistance each time a certain task is to be performed or particular conditions are encountered. The user's current affective state and the importance of the current task given the context of use could be exploited to time the interrupts conveniently. For example, if the user is hurriedly writing e-mails, interrupting him to correct a syntax mistake can be postponed till the moment he tries to send the e-mail.

When, in which part, and in what way should the system be adjusted? The sensed user's affective state could be exploited to time the adjustment of the system, the information about the user's current task might form the target of the adjustment, and the complete information about the sensed context could be used to determine the adjustment properly. For instance, suppose that the user always browses through a particular application in the same way in order to come to a specific window and displays irritation each time the system starts the pertinent application by displaying its very first window. In that case, a proper adjustment might be to mark the window where the user commonly stops browsing and to start the pertinent application with that specific window (browsing through the preceding windows of the application does not have to be apparent to the user). Yet, suppose that the user always becomes frustrated if a certain person enters the office. In that case, no adjustment should be

made since the user's affective state is not caused by HCI but by an external (and for HCI) irrelevant event.

HCI design breakthroughs are necessary if the computing technology is to achieve the ultimate goal of being universally usable [43]. However, many research problems initially thought to be intractable have been proven manageable (e.g., sensing and detecting human interactive cues [33], affect-sensitive interpretation of those [34], [36], and context sensing in general [29], [33]). This initiated numerous efforts to design and develop multimodal and context-sensitive interfaces that are flexible and compatible with users' abilities and relevant work practices [30], [31].

1.2 The State of the Art

Several extensive survey and position papers have been published on vision-based [40], [52], multimodal [30], [31], [19], affective [24], [17], and context-sensitive interfaces [11], [29], [33]. Virtually all of these articles agree that approaching the naturalness of human-human interaction plays a central role in the future of HCI and that this objective can be approached by designing adaptive HCI systems that are affective, context-aware, and multimodal. However, many of these articles mention that the main application of this new technology is quickly changing from user interface models in which one user is sitting in front of the computer, to something else like ambient interface models in which multimodal multi-party interactions are observed and interpreted. We argue here that this statement may be very misleading. When the main application domain in a certain field has changed from *A* to *B*, this may imply (and it usually does imply) that problems in *A* have been researched, that they have been solved, and that the research has moved on to tackle other problems. In turn, the statement in question may imply that the realization of adaptive interfaces based on affective multimodal interaction models (AMM-HCI) can be considered a solved problem for single-user office scenarios. However, an extensive research of the large body of the related literature did not confirm this.

Only few works aimed at adaptive, affective, multimodal interfaces for single-user office scenarios have been reported up to date. The majority of past work in the field relate to multimodal, non-affective interaction with the user [30], [31]. Integration of multiple natural-interaction modalities such as speech, lip reading, hand gestures, eye-tracking, and writing into human-computer interfaces has long been viewed as a means for increasing the naturalness and, in turn, ease of use [3], [53], [54]. The research in this part of the field is still very active and the majority of the current work focuses on speak-and-point multimodal interaction (where pointing is conducted by either pen or hand gesture) and aimes either at support of crisis management [27], [46], or at development of personal widgets [30].

A rather large body of research can also be found in the field of human affect sensing [34], [36], [33]. Most of these works are single modal, based either on facial or on vocal affect analysis. Recently, few works have been also proposed that combine two modalities into a single system for human affect analysis. The majority of these efforts aim mainly at combining facial and vocal expressions for affect recognition (e.g., [5], [55]), although some tentative attempts on combining facial expressions and body postures have been reported as well [14], [20]. In the same way that these methods do not tackle the problem of how the sensed user's affect can be

incorporated into the HCI, there is a large body of research that focuses on the role of human affect in HCI while assuming that user's affective states have been already sensed by the computer [7], [44]. Efforts that integrate these two detached research streams, and represent the related work to the one presented in this paper, are rare. Several works have been reported on developing adaptive interfaces that are based on sensing the user's affective states. These methods are usually single modal, based either on facial affect recognition (e.g., [2]) or on physiological affective reaction recognition (e.g., [18], [42]). To the best of our knowledge, the only exceptions from this rule are the single-user AMM-HCI systems proposed by Lisetti & Nasoz [23] and by Duric et al. [11]. The former combines facial expression and physiological signals to recognize the user's emotion (fear, anger, sadness, frustration, and neutral) and then to adapt an animated interface agent to mirror the user's emotion. The latter is a much more elaborate system, aimed at real-world office scenarios. It combines lower arm movements, gaze direction, eyes and mouth shapes, as well as the kinematics of mouse movements to encode the user's affective and cognitive states (confusion, fatigue, stress, lapses of attention, and misunderstanding of procedures). It then applies a model of embodied cognition, which can be seen as a detailed mapping between the user's affective states and the types of interface adaptations that the system supports, to adapt the interface in a reactive or a proactive manner to the user's affective feedback. The main drawback of this system is that it is not user-profiled, while different users may have different preferences for both the input modes per type of action and the type of interface adaptation. Another drawback is the employed model of embodied cognition, which is rigid and clumsy as it stores all possible combinations of inputs and outputs that are difficult to unlearn and reformat. Finally, as the system was not tested and evaluated, conclusions about its robustness, flexibility, adaptability, and overall usability cannot be drawn.

We believe that the main reason for the lack of research on single-user AMM-HCI is a twofold. First, the highly-probable misconception that the problem in question has been solved could cause the lack of interest by researchers and research sponsors. Second, it seems that the vast majority of researchers treat the problem of adaptive, affective, multimodal interfaces as a set of detached problems in human affect sensing, human interactive signals processing, human-human interaction modeling for HCI, and computer-human interface design. In this paper, we treat this problem as one complex problem rather than a set of detached problems and we propose a single-user AMM-HCI system similar to that proposed by Duric et al. [11]. A difference between the two systems is that ours uses a dynamic case base, an incrementally self-organizing event-content-addressable memory that reflects user preferences and allows easy reformatting each time the user wishes so. In addition, our system, which we call Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how) and, in turn, is user- and context-profiled.

The state of the art in context-aware applications ensues from two streams of research: the one on context sensing [29], [33], which focuses on sensor-signal processing (audio, visual, tactile), and the other on context modelling [10], which focuses on specifying procedures and requirements for all pieces of context information that will be followed by a context-aware application. Gaze-X integrates those two detached poles of the research. It uses a face recognition system to answer

who the user is and to retract his or her profile (i.e., user-profiled case base). It employs an eye-tracking system and a speech recognizer in combination with event handling of standard HCI events like mouse movements, keystrokes, and active software identification to answer *what* is the current task of the user. In addition to these input modalities, a system that recognizes prototypic facial expressions of six basic emotions (anger, fear, happiness, surprise, sadness, and disgust) is used to answer the *how* context question. To answer the *when* context question, we simply keep a log of the time and the cost (in time) of HCI events associated with various input modalities. To answer the *why* context question, which is the most complex context question, we use case-based reasoning that enables evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Based upon the conducted evaluation, Gaze-X executes the most appropriate user-supportive action. The system was initially proposed by the authors in [25].

1.3 Organization of the Paper

The paper is organized as follows. Section 2 gives an overview of the Gaze-X architecture. Section 3 presents the system's input modalities. The utilized case-based reasoning is explained in detail in section 4. Adaptive and user-supportive actions of the interface are discussed in section 5. The Graphical User Interface (GUI) of the Gaze-X is presented in section 6. The usability study that we carried out is discussed in section 7. Section 8 concludes the paper.

2 System Architecture

The outline of the system is illustrated in Fig. 1. The main modules are the multimodal input module, the reasoning module, and the feedback module. The user of Gaze-X experiences an adaptive interface, which changes as a function of the currently sensed context. This function is represented by the cases of the utilized dynamic case base, having the system's and the user's state as the input (represented in terms of exhibited multimodal interactive actions and cues) and the adaptive and user-supportive changes in the interaction as the output. The fact that the changes in the interaction do not have to occur in each time instance (e.g., if the user's preference is to remain undisturbed while working with a certain application), triggering of the feedback module is optional and illustrated using a dashed line.

Gaze-X has been implemented as an agent-based system. The main reason for doing so is to support concepts of *concurrency* (which allows sub-systems to operate independently and yet at the same time), *modularity/scalability* (which allows easy upgrade through inclusion of additional sub-systems), *persistency* (which ensures robust performance by saving intermediate settings), and *mobility* (which enables transport of agents to another host). We used Fleeble Agent Framework [32] to develop the Gaze-X. Fleeble can be seen as a common programming interface defining the behavior of all the agents build with the framework. The main characteristics of Fleeble and Fleeble-based multi-agent systems (MAS) can be summarized as follows (for details see [32]).

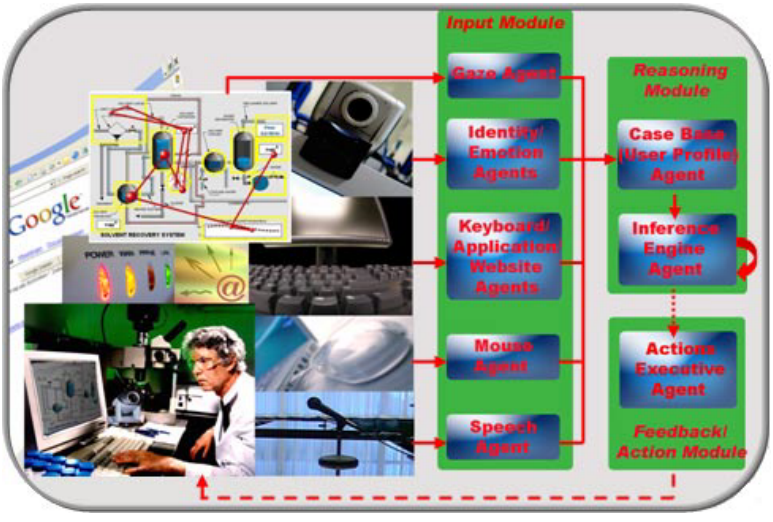


Fig. 1. Overview of the Gaze-X system: the input modalities, modules, and agents

Fleeble enables easy development of agents and MAS. The framework can instantiate and configure an agent and then start it up in a separate thread. The agent is autonym, although it is running in the framework's processing space. An agent can also instruct the framework to start up another agent. The framework keeps track of all agents and their parent agents. So, a single agent can be created which starts up the appropriate agents for each MAS. This *kickoff* agent is then the parent agent of all agents that form the MAS in question.

Fleeble supports simple processing of events coming from the outside world and other agents. Fleeble offers a message distribution system for communication between agents that is based on a Publish/ Subscribe system, which is centered on the concept of a *channel*. Channels are named entities that allow a single message to be delivered to any number of agents. An agent informs Fleeble that it is interested in events pertaining to a specific channel (i.e., it subscribes to that channel). An agent can ask Fleeble to deliver a message to a channel (i.e. it publishes to the channel in question). Fleeble creates a "handler" thread for each agent that has subscribed to the channel in question. All handler threads are started at the same time and deliver the message to the subscribed agents. Hence, e.g., the user's facial cues can be simultaneously processed by both the Identity Agent and the Emotion Agent (Fig. 1).

Fleeble supports the concept of concurrency needed to allow agents to operate independently and yet at the same time. This has been achieved by starting each agent in a separate thread, allowing it to access the delivery system described above at its own convenience.

Fleeble supports data and state persistence. Fleeble agents can instruct the framework to store values referenced by a key. The framework stores this (key, value) pair and allows access to it at any time, even when the execution of the framework has ceased in the meantime. State persistency allows the user to shut down a single agent (or MAS) and to restore it from the point where it was suspended later on, even

when the PC has been shut down in the meantime. This makes Fleeble-based MAS very robust.

Fleeble supports the concepts of distribution and mobility. It establishes socket-to-socket connections between frameworks residing on different computers and manages these connections, i.e., creates and closes them as needed. Connections can be used to transport agents to another host, allowing agents to physically move. The process of being moved to another host can be started either by the agent or by any parent agent.

3 Input Modalities

The front end of Gaze-X consists of the multimodal input module, which processes images of the user's face, gaze direction, speech, and actions done while interacting with the computer including the mouse movements and the keystrokes.

The choice of input modes was largely influenced by findings in cognitive science and HCI-related research on multisensory perception and multimodal coordination. These findings reveal that people have strong preference to interact multimodally when engaged in spatial location commands, whereas they tend to interact unimodally when engaged in actions that lack a spatial component. More specifically, it seems that people prefer speak-and-point interaction in visual-spatial domains like Graphical User Interfaces [15], whereas they prefer speech input alone for issuing commands for actions and for describing objects, events, and out-of-view objects [6]. Although researchers discarded the idea of using gaze fixations as a mouse replacement, especially for extended periods, gaze direction is exploited nowadays as a reliable index of the user's interest [25]. Finally, the face is our primary means to identify other members of the species, and to communicate and understand somebody's affective state and intentions on the basis of shown facial expressions [21]. Hence, integrating face, facial expression, speech, mouse-pointing, and gaze direction input modes into the Gaze-X, promised a suitably complementary modality combination that could support the intended, adaptive, context-sensitive (user-, task-, and affect-sensitive) interaction with the user.

To process the images of the user's face acquired by a standard web-cam, we use a commercially available the Face Reader system for face and facial affect recognition produced by Vicar Vision.[1] This system operates as follows. It detects candidate face regions in the input scene by comparing image regions to a number of prototype faces. These prototypes are representative of a large database of human faces. An Active Appearance Model (AAM) [8] is then fitted to the detected face region (Fig. 2). The variations in the shape and texture of the AAM, caused by fitting the AAM to the detected face, serve as a unique identifier of the individual (identifiers for different users are stored in a database). Once the user is identified, AAM fitting is employed to detect the affective state of the user. The variations in the shape and texture of the AAM, caused by fitting the AAM to the user's face in the current frame, are fed to a neural network trained to recognize the prototypic facial expressions of six basic emotions (surprise, fear, anger, happiness, sadness, and disgust) [22], defined in classic psychological studies on human emotions [21]. The

[1] Vicar Vision BV, 2004. http://www.vicarvision.nl

Fig. 2. Person identification of the Face Reader system [22]. From left to right: input image, automatic placement of facial landmarks, initial positioning of the AAM, best-fitting face.

output layer of the neural network consists of 7 nodes (one for each emotion and one for the neutral state), each of which outputs a continuous value. This enables detection of blends of emotions (e.g., surprise and happiness in an expression of delight) and low-intensity emotions (e.g., a frown is recognized as a low intensity of anger).

To detect the user's gaze direction, we employ a commercial system for remote eye tracking produced by SMI GmbH.[2] The utilized iView X remote eye-tracking system consists of two main components, the infrared pan tilt camera and the software used for the calibration and eye tracking. To determine the direction of the gaze, the system employs the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil. This direction determines the point on the screen, which is the focus of the user's attention (Fig. 3). As a cursor can be displayed at this point, eye-tracking can be employed to free the user from the constraints of the mouse and, in combination with spoken commands, to allow a hands-free interaction with the computer.

A variety of speech-recognition systems are now available either as commercial products or as open source. Although some novel approaches to speech recognition has been proposed recently [9], most of the existing methods utilize acoustic information of speech contained in frame-by-frame spectral envelopes which are statistically classified by Hidden Markov Models. Gaze-X utilize Sphinx 4[3] speech recognizer, which is a Java-implemented speech recognizer that recognizes a predefined set of words (vocabulary) based on acoustic features and a HMM architecture.

To monitor standard HCI events including keystrokes, currently active software and currently visited web-site, we utilize the Best Free Keylogger software.[4] It monitors website visiting/blocking, e-mail visiting, keystrokes, and application activity, and writes these data into log files. We also monitor and log the locations of the mouse cursor. In the case that the user's preference is to use eye-tracker as an alternative for the mouse, the system does not log the mouse movements.

[2] SensoMotoric Instruments GmbH, 2002. http://www.smi.de
[3] Sphinx-4, 2004. http://cmusphinx.sourceforge.net/sphinx4
[4] Best Free Keylogger, 2006. http://sourceforge.net/projects/bfk/

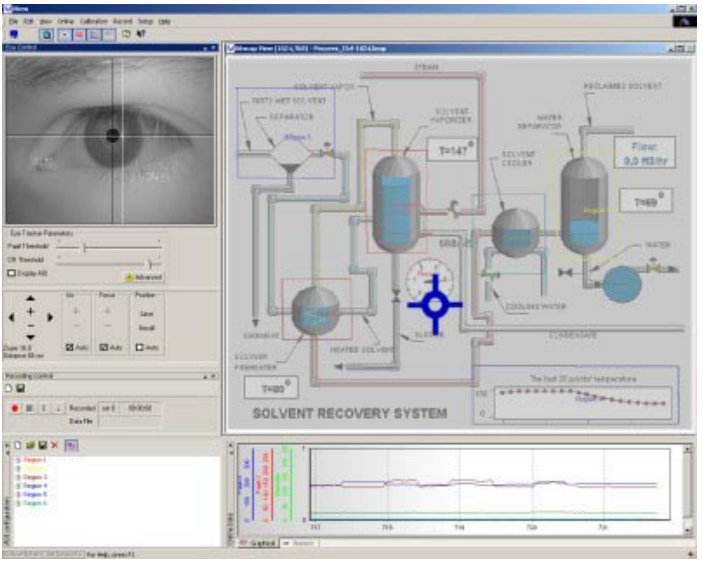


Fig. 3. Determining the user's gaze direction (focus of attention) by iView X eye-tracker

The user's identity, his or her displayed affective state, current gaze direction, and uttered words, delimit the current user's state. The HCI events including the mouse movements, keystrokes, and the currently active software delimit the current system's state. These two states form further the input to the reasoning module.

4 Case-Based Reasoning

Since the Gaze-X can have different users, each of which can be using different applications in his or her daily work with the PC, while showing emotions and a variety of interactive patterns using the standard and the natural interactive modalities that the Gaze-X supports, the mapping of the system's multimodal input onto a large number of adaptive and user-supportive changes in interface in a user-profiled manner is an extremely complex problem. To tackle this problem, one can apply either eager or lazy learning methods. Eager learning methods such as neural networks extract as much information as possible from training data and construct a general approximation of the target function. Lazy learning methods such as case-based reasoning store the presented data and generalizing beyond these data is postponed until an explicit request is made. When a query instance is encountered, similar related instances are retrieved from the memory and used to classify the new instance. Hence, lazy methods have the option of selecting a different local approximation of the target function for each presented query instance [1]. Eager methods using the same hypothesis space are more restricted since they must choose their approximation of the target function before query instances are observed. In turn, lazy methods are usually more appropriate for complex and incomplete problem domains than eager methods, which replace the training data with abstractions obtained by generalization

and which, in turn, require excessive amount of training data. Hence, we chose to implement the inference engine of the Gaze-X as the case-based reasoning on the content of a dynamic memory. The memory is dynamic in the sense that, besides generating user-supportive feedback by analogy to that provided to the user in similar situations "experienced" in the past, it is able to unlearn feedback actions that the user liked once but now tends to dislike and to learn new feedback actions according to the instructions of the user, thereby increasing its expertise in user-profiled, user-supportive, adaptive user-computer interaction.

The utilized dynamic memory of experiences is based on Schank's theory of functional organization of human memory of experiences [45]. According to this theory, for a certain event to remind one spontaneously of another, both events must be represented within the same dynamic chunking memory structure, which organizes the experienced events according to their thematic similarities. Both events must be indexed further by a similar explanatory theme that has sufficient salience in the person's experience to have merited such indexing in the past. Indexing, in fact, defines the scheme for retrieval of events from the memory. The best indexing is the one that will return events most relevant for the event just encountered.

In the case of Gaze-X memory of experiences, each event is one or more micro-events, each of which is an interactive cue (a part of the system's multimodal input) displayed by the user while interacting with the computer. Micro-events that trigger a specific user-supportive action are grouped within the same dynamic memory chunk. The indexes associated with each chunk comprise individual micro-events and their combinations that are most characteristic for the user-supportive action in question. Finally, micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of times the user was satisfied when the related user-supportive action was executed as the given micro-event occurred, the higher the hierarchical position of that micro-event within the given chunk. Certain user-supportive action can be preferred by the user for both different software applications that he or she is usually using and different affective states that he or she is displaying. Hence, to optimize the search through the case base, affective states and currently active software are not treated as other micro-events but are used as containers of various memory chunks. More specifically, memory chunks representing user-supportive actions that are triggered when the user is displaying a certain emotion are grouped in a super class representing that emotion. Hence, each chunk may contain a pointer to an emotion-identifying super class. Similarly, each chunk may contain a pointer to an active-software-identifying super class. A schematic representation of Gaze-X case base organization is given in Fig. 4.

To decide which user-supportive action is to be executed (if any) given an input set of interactive cues displayed by the user, the following steps are taken:

- Search the dynamic memory for similar cases based on the input set of observed interactive cues, retrieve them, and trigger the user-supportive action suggested by the retrieved cases.
- If the user is satisfied with the executed action, store the case in the dynamic memory and increase its typicality. If the user is not satisfied with the executed action, adapt the dynamic memory by decreasing the typicality of the just executed action.

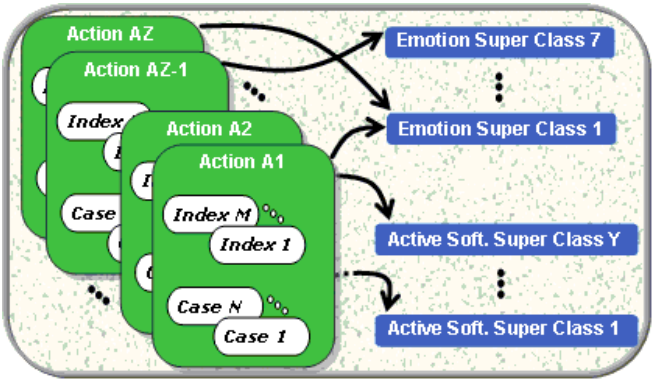


Fig. 4. Schematic organization of the case base utilized by the Gaze-X. Each user has his or her own personal case base (user profile).

The simplest form of retrieval is to apply the first nearest neighbor algorithm, that is, to match all cases of the case base and return a single best match. This method is usually to slow. A pre-selection of cases is therefore usually made based on the indexing structure of the utilized case base. Our retrieval algorithm employs a pre-selection of cases that is based on the clustered organization of the case base (super classes and memory chunks), the indexing structure of the memory, and the hierarchical organization of cases within the memory chunks according to their typicality.

Gaze-X can run in two modes, an unsupervised and a supervised mode. In the unsupervised mode, the affective state of the user is used to decide on his or her satisfaction with the executed action. If a happy or a neutral expression is displayed, Gaze-X assumes that the user is satisfied. Otherwise, if the user emotes negatively, Gaze-X assumes that he or she is dissatisfied. In the supervised mode, the user explicitly confirms that an action of his preference has been executed. If the user is not satisfied with the executed action, the dynamic memory is adapted. In the unsupervised mode, the typicality of the relevant case is decreased. In the supervised mode, the typicality of the relevant case is decreased and the user may provide further feedback on the action of his/her preference that should be executed instead.

5 Interaction Adaptation

The final processing step of Gaze-X is to adapt the user-computer interaction based on current needs and preferences of the user and according to adaptive and user-supportive changes suggested by the system's dynamic memory of experiences. General types of interface adaptations supported by Gaze-X include the following.

Help provision – Examples include the following. When the open-file-dialogue is open for a long time, help can be provided by highlighting the file names that were opened in the past in combination with the currently open files. Alternatively, desktop search application can be started. If the user selected a column in a table, and (s)he is

scrutinizing the menu bar for a long time, help can be given by highlighting table-related menu options. Alternatively, help-menu option can be highlighted.

Addition/removal of automation of tasks – Examples include automatic opening of all windows of an application that the user opens each time he or she starts up the application in question, automatic error correction, automatic blockage of websites that are similar to already blocked sites, and removal of such an automation if the user disapproves of it.

Changing information presentation – Automatic selection of most relevant features/options to be displayed given the user's current task, automatic font size increase/decrease according to the user's preferences, usage of eye tracking and speech as an alternative to mouse movements, automatic sound play, etc, are typical examples of this type of interface adaptation.

Gaze-X carries out interface changes in a rather conservative way. More specifically, when operating in the unsupervised mode, it executes adaptive and user-supportive actions one at the time and in a rather slow pace. The underlying philosophy is not to make an ever-changing interface a source of user's frustration in itself. In order to allow the user some time to get accustomed to the idea of a self-adaptive interface and to initialize the system using a (small) set of interface changes that the user considers helpful, Gaze-X is initially set to operate in the supervised mode. As soon as the user considers the system profiled enough, he or she can set the system to operate in the unsupervised mode.

6 GUI of the System

Gaze-X has a simple, easy to understand GUI. The goal was to develop a direct-manipulation GUI in which WYSIWYG (what you see is what you get) would be the guiding principle. A simple self-explicatory GUI was developed that is easy to understand and use. As can be seen in Fig. 5, the main window of Gaze-X GUI visualizes who the current user is (i.e., whose user-profiled case base is currently used), allows loading, creation, and adaptation of the profiles, and enables initiation of adaptive interaction mode supported by the system. Gaze-X automatically loads the user's profile for a known user based on the output of the Face Reader system as explained in section 3. However, an option to load a profile manually, using a valid username and password, also exists. Besides, since Gaze-X is developed as Fleeble-based MAS, the agents that constitute the system including the reasoning agent (representing the profile of the user) are mobile and can be moved to another host. Hence, users' profiles can be transmitted as needed to any computer where Gaze-X is installed.

When installing Gaze-X, a directory needs to be found where the main system and the supporting systems and sensors (Fleeble, the FaceReader face and facial expression detector, the iView-X remote eye tracker, a web cam, and a microphone) can be installed with proper access rights. Also a specific version of Java Runtime Environment (version 1.5 or later) needs to be properly installed before Gaze-X can run. To make this process as easy as possible for the user, a setup wizard has been implemented. It executes automatically many of the required steps to set up Gaze-X and leads the user through the rest of the required steps to ensure that the required hardware and software are properly setup.

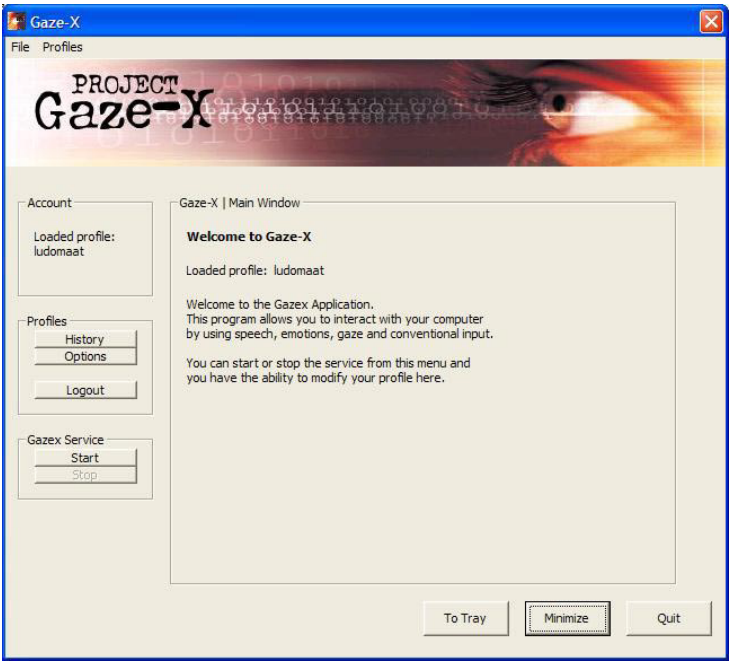


Fig. 5. The main window of the Graphical User Interface of the Gaze-X

Any new way of thinking about programming takes some getting used to, and Gaze-X is no exception to this axiom. To aid users in working out questions that they may have, Gaze-X provides a tutorial. It shows the basic functionalities and usage of Gaze-X in a step by step demonstration. The tutorial can be started by means of 'To Try' button illustrated in Fig. 5.

7 Usability Study

To make a preliminary assessment of effectiveness, usability, usefulness, affective quality, and ethical issues relevant to Gaze-X, we conducted a small evaluation study with the help of six participants. The participants were 18 to 61 years old, 33% female, 50% Caucasian, 33% Asian, 17% African, 17% expert, 50% intermediate, and 33% novel computer users. We asked them to install Gaze-X first and then to use it as demonstrated by the tutorial integrated into the system. For each session with another user, we used either a Linux or a Windows machine from which the Gaze-X was removed. We did not require the users to use the Gaze-X for a certain period of time, each of them was engaged in exploring Gaze-X for as long as he or she wanted. We also did not require from the participants to work with specific software applications. Note, however, that the software installed on the machines we used for the experiments had an Internet browser, an e-mail handler, a text editor, Adobe Reader, and a number of multimedia handlers such as music and movie players.

We used a custom-made questionnaire to elicit users' attitudes towards the Gaze-X-supported interaction with the computer. The questionnaire includes questions soliciting users' attitudes toward:

- the effectiveness of the HCI design (i.e., whether the interaction with computer is more natural than it is the case with standard HCI designs and whether it is robust enough, [43]),
- the usability of Gaze-X (i.e., whether technological variety and user diversity are supported and whether gaps in the user's computer knowledge play an important role, [48]),
- the usefulness of the system (i.e., whether the utility of system's functionalities and the utility of interface adaptation supported by the system are obvious to the users, whether they will choose to use the system if it was publicly available),
- the ethical issues related to the HCI design (i.e., do users feel uncomfortable under the scrutiny of machines that monitor their affective states, work- and interaction patterns, will users adapt soon to emote just to make the computer do something, [24]),
- the affective quality of the HCI design (i.e., whether the GUI of the Gaze-X is aesthetically qualitative in terms of orderly, clear, and creative design, [56]).

The utilized questionnaire also invites participants' suggestions on how to improve the system in any of the aspects mentioned above. It employs a 5-point Likert scale ranging from strongly disagree (1), via neutral (3), to strongly agree (5). 'I do not know' is also a possible answer. The main points of the obtained survey results are listed in Table 1.

Table 1. Users' satisfaction with effectiveness, usability, usefulness, affective quality, and ethical issues relevant to Gaze-X. The percentages in the table show the percentage of *agree* and *strongly agree* answers. EU stands for expert users (total: 17%) and NU stands for intermediate and novel users (total: 83%). × indicates an 'I do not know' answer.

Survey question	EU	NU
The interaction is more natural than in standard HCI	0%	100%
The interaction is robust enough	0%	66%
Gaze-X supports technological variety	×	100%
Gaze-X supports users of different age, skills, culture…	100%	100%
Gaze-X is easy to use even if users lack IT knowledge	100%	100%
Face identification is a useful functionality	100%	100%
Monitoring affective states is a useful functionality	0%	66%
Having mobile user profiles is useful	100%	100%
Having multimodal interaction is useful	0%	66%
Adaptive and user-supportive interface is useful	0%	100%
Gaze-X makes the interaction with computer easier	0%	100%
I do not mind to be observed by the camera	100%	100%
I like the aesthetics of the Gaze-X GUI design	100%	100%
I would use Gaze-X if it was publicly available	0%	66%

All participants seem to agree on the usability, affective quality, and ethical issues relevant to Gaze-X. The perceived usability is directly related to the following properties of Gaze-X: it runs on various platforms like Linux, Windows, and Mac Os X (as a result of being Java-implemented), it accommodates users of different age, gender, culture, computing skills and knowledge, and it bridges the gap between what the user knows about the system and HCI in general and what he or she needs to know (it provides a setup wizard and a tutorial that shows how to use Gaze-X in a step by step demonstration). The perceived affective quality of Gaze-X is directly tied to aesthetics qualities of the Gaze-X GUI: (i) it has an orderly and clear design in accordance to the rules advocated by usability experts, and (ii) it reveals designers' creativity, originality, and the ability to break design conventions manifested by, for example, multimodal design, affective design, tutorial demonstration, etc. These findings are consistent with research finding of Zhang and Li [56], who suggested that the perceived affective quality of a software product is directly tied to aesthetics qualities of that product and who argued that affectively qualitative products have significantly larger chance to be widely accepted technology. The perceived ethical issues relevant to Gaze-X were somewhat surprising as it seems that the users have no problem with being continuously observed by a web cam. Standard concern that the user's behavioral and affective patterns could be used to mind-read and manipulate him or her was not mentioned a single time. However, all participants in our study stressed the importance of privacy and asked about measures taken to prevent hacking and intrusion. Ultimately, if users have control about whether, when, and with whom they will share their private information such as the observations of their behavioral patterns while interacting with the PC (as stored in their personal profile), fears from big-brother-is-watching-you scenarios vanish.

However, as can be seen from Table 1, the question that remains is whether or not an individual user values the capability of Gaze-X to be aware of his or her behavioral and affective patterns. Our study suggests that there might be a very large difference between the expert and non expert computer users when it comes to this question. While all users agreed that having a face-identification-based access to the system is a very useful functionality, expert users found other functionalities of Gaze-X less appealing and in some instances irritating (e.g., popup widows used in supervised operating mode to ask the feedback about the user's preferences). On the other hand, less experienced computer users perceived Gaze-X as very useful since it provides support when needed, in a way needed and preferred, making the interaction with the computer more efficient and easier (e.g., less time was spent on searching various functionalities and undoing erroneous actions). In turn, it seems that Gaze-X is very suitable for novel and intermediate computer users but is much less so for experienced users. As suggested by the expert computer user who participated in our study, much more sophisticated user support should be provided to an experienced user than it is currently the case. For example, support should be provided only when a new application is installed and used for the first couple of times, no support should be provided for long installed software applications. Note, however, that only one expert computer user has participated in the present usability study. A much more elaborate survey must be conducted with experienced users if some firm conclusions are to be made about the ways to make Gaze-X useful and appealing to experienced computer users.

Finally, all participants remarked that the robustness of the system can be improved. Two issues are of importance here. The first is the sensitivity of the Face Reader system, used for face and facial expression analysis, to changes in lighting conditions. The second is the sensitivity of the iView-X remote eye tracker to changes in the user's position. More specifically, the user is expected to remain in front of the computer and is not allowed to shift his or her position in any direction for more than 30 cm. Otherwise, the iView-X system should be recalibrated before it can be used again. These findings indicate that in the future more robust systems for facial expression analysis and eye tracking should be considered for inclusion in the Gaze-X.

From other remarks mentioned by the participants in the present study, arguably the most important one relates to the choice of the affective states to be tagged by Gaze-X. Most of participants said that they may experience confusion, frustration, understanding, tiredness, and satisfaction while interacting with the computer. However, the currently employed Face Reader system recognizes only facial expressions of six basic emotions including disgust, fear, and sadness, for which the participants in our study said that they are not likely to be experienced in a HCI setting like office scenarios. This indicates that in the future we should employ either automatic analyzers of attitudinal and non-basic affective states like attentiveness [12] and fatigue [13], or systems for user-profiled interpretation of facial expressions [35].

8 Conclusions

In this paper we proposed one of the first systems for adaptive, affective, multimodal human-computer interaction in standard office scenarios. Our system, Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how) and, in turn, is user- and context-profiled. The user of Gaze-X experiences an adaptive interface, which changes as a function of the currently sensed context. This function is represented by the cases of the utilized dynamic case base, having the system's and the user's state as the input (represented in terms of exhibited multimodal interactive actions and cues) and the adaptive and user-supportive changes in the interaction as the output. A usability study conducted in an office scenario with the help of six users indicates that Gaze-X is perceived as effective, easy to use, and useful by novice and less experienced users and as usable and affectively qualitative by all participants in the present study. In turn, Gaze-X seems to be very suitable for novel and intermediate computer users but is much less so for experienced users. As only one experienced user has participated in the present usability study, a much more elaborate survey must be conducted with experienced users if some firm conclusions are to be made about the ways to make Gaze-X useful and appealing to this group of users. Ultimately, as the majority of current software products are still designed for experienced frequent users and designing for a broad audience of unskilled users is still seen as a far greater challenge [48], we are very glad and proud that Gaze-X was perceived as useful and easy to use by novice and less experienced users, who in general still experience computing technology as too difficult to use. Except of a more elaborate usability study with a large number of various users, the focus of our future research will also be aimed at enabling the system to tag attitudinal and non-basic affective states such as confusion, stress, fatigue and satisfaction.

Acknowledgments

The authors would like to thank Marten den Uyl of Vicar Vision BV for providing the Face Reader system. The work has been conducted at Delft University of Technology. The work of M. Pantic has been supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202. The cooperation with Vicar Vision BV has been carried out in the scope of the Dutch BSIK-MultimediaN-N2 project on Interaction.

References

1. Bartsch-Sporl, B., Lez, M., Hubner, A.: Case-based reasoning – survey and future directions. Lecture Notes in Artificial Intelligence, Vol. 1570 (1999) 67-89
2. Bianchi-Berthouze, N., Lisetti, C.L.: Modeling multimodal expression of user's affective subjective experience. User Modeling and User-Adapted Interaction, Vol. 12, No. 1 (2002) 49-84
3. Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. Computer Graphics, Vol. 14, No. 3 (1980) 262-270 (Proc. ACM SIGGRAPH'80)
4. Browne, D., Norman, M., Riches, D.: Why Build Adaptive Interfaces? In: Browne, D., Totterdell, P., Norman, M. (eds.): Adaptive User Interfaces. Academic Press, London, UK (1990) 15-57
5. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling Naturalistic Affective States via Facial and Vocal Expressions Recognition. Proc. Int'l Conf. Multimodal Interfaces (2006) 146-154
6. Cohen, P., Oviatt, S.L.: The role of voice input for human-machine communication. Proc. National Academy of Sciences, Vol. 92 (1995) 9921-9927
7. Conati, C.: Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence, Vol. 16, No. 7-8 (2002) 555-575
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, No. 6 (2001) 681-685
9. Deng, B.L., Huang, X.: Challenges in adopting speech recognition. Communications of the ACM, Vol. 47, No. 1 (2004) 69-75
10. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. J. Human-Computer Interaction, Vol. 16, No. 2-4 (2001) 97-166
11. Duric, Z., Gray, W.D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M.J., Schunn, C., Wechsler, H.: Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proceedings of the IEEE, Vol. 90, No. 7 (2002) 1272-1289
12. El Kaliouby, R., Robinson, P.: Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. Proc. Int'l Conf. Computer Vision & Pattern Recognition, Vol. 3 (2004) 154-
13. Gu, H., Ji, Q.: An automated face reader for fatigue detection. Proc. Int'l Conf. Face & Gesture Recognition (2006) 111-116
14. Gunes, H., Piccardi, M.: Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. Proc. Int'l Conf. Systems, Man and Cybernetics (2005) 3437- 3443
15. Hauptmann, A.G.: Speech and gestures for graphic image manipulation. Proc. ACM Int'l Conf. Human Factors in Computing Systems (1989) 241-245

16. Hoffman, D.L., Novak, T.P., Venkatesh, A.: Has the Internet become indispensable? Communications of the ACM, Vol. 47, No. 7 (2004) 37-42
17. Hudlicka, E.: To feel or not to feel: The role of affect in human-computer interaction. Int'l J. Human-Computer Studies, Vol. 59, No. 1-2 (2003) 1-32
18. Hudlicka, E., McNeese, M.D.: Assessment of user affective/belief states for interface adaptation. User Modeling & User-Adapted Interaction, Vol. 12, No. 1 (2002) 1-47
19. Jaimes, A., Sebe, N.: Multimodal human computer interaction: A survey. Proc. IEEE ICCV Int'l Workshop on HCI in conjunction with Int'l Conf. Computer Vision (2005)
20. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. Proc. ACM Int'l Conf. Multimedia (2005) 677-682
21. Keltner, D., Ekman, P.: Facial expression of emotion. In: Lewis, M., Haviland-Jones, J.M. (eds.): Handbook of Emotions. The Guilford Press, New York (2000) 236-249
22. van Kuilenburg, H., Wiering, M., den Uyl, M.: A model-based method for automatic facial expression recognition. Lecture Notes in Artificial Intelligence, Vol. 3720 (2005) 194-205
23. Lisetti, C.L., Nasoz, F.: MAUI: A multimodal affective user interface. Proc. Int'l Conf. Multimedia (2002) 161-170
24. Lisetti, C.L., Schiano, D.J.: Automatic facial expression interpretation: Where human-computer interaction, AI and cognitive science intersect. Pragmatics and Cognition, Vol. 8, No. 1 (2000) 185-235
25. Maat, L., Pantic, M.: Gaze-X: Adaptive affective multimodal interface for single-user office scenarios. Proc. Int'l Conf. Multimodal Interfaces (2006) 171-178
26. Maglio, P.P., Barrett, R., Campbell, C.S., Selker, T.T.: SUITOR: An attentive information system. Proc. Int'l Cof. Intelligent User Interfaces (2000) 169-176
27. Marsic, I., Medl, A., Flanagan, J.: Natural communication with information systems. Proceedings of the IEEE, Vol. 88, No. 8 (2000) 1354-1366
28. Nielsen, J.: Multimedia and hypertext: The Internet and beyond. Academic Press, Cambridge, USA (1995)
29. Nock, H.J., Iyengar, G., Neti, C.: Multimodal processing by finding common cause. Communications of the ACM, Vol. 47, No. 1 (2004) 51-56
30. Oviatt, S. User-centred modelling and evaluation of multimodal interfaces. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1457-1468
31. Oviatt, S., Darrell, T., Flickner, M.: Multimodal Interfaces that Flex, Adapt, and Persist. Communications of the ACM, Vol. 47, No. 1 (2004) 30-33
32. Pantic, M., Grootjans, R.J., Zwitserloot, R.: Teaching Ad-hoc Networks using a Simple Agent Framework. Proc. Int'l Conf. Information Technology Based Higher Education and Training (2005) 6-11
33. Pantic, M., Pentland, A., Nijholt, A., Huang, T.S.: Human computing and machine understanding of human behaviour: A survey. Proc. Int'l Conf. Multimodal Interfaces (2006) 239-248
34. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1370-1390
35. Pantic, M., Rothkrantz, L.J.M.: Case-based reasoning for user-profiled recognition of emotions from face images. Proc. Int'l Conf. Multimedia and Expo (2004) 391-394
36. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S.: Affective multimodal human-computer interaction. Proc. ACM Int'l Conf. Multimedia (2005) 669-676
37. Pentland, A.: Looking at people: Sensing for ubiquitous and wearable computing. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 1 (2000) 107-119
38. Pentland, A.: Perceptual intelligence. Communications of the ACM, Vol. 43, No. 3 (2000) 35-44

39. Picard, R.W.: Affective Computing. The MIT Press, Cambridge, USA (1997)
40. Porta, M.: Vision-based user interfaces: methods and applications. Int'l J. Human-Computer Studies, Vol. 57, No. 1 (2002) 27-73
41. Preece, J., Schneiderman, B.: Survival of the fittest: Evolution of multimedia user interfaces. ACM Computing Surveys, Vol. 27, No. 4 (1995) 557-559
42. Prendinger, H., Ishizuka, M.: The empathic companion: A character-based interface that addresses users' affective states. Applied Artificial Intelligence, Vol. 19, No. 3-4 (2005) 267-285
43. Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T.V., Stanney, K.M., Su, H., Wang, Q.Y.: Guidelines for multimodal user interface design. Communications of the ACM, Vol. 47, No. 1 (2004) 57-59
44. Ruttkay, Z., Pelachaud, C. (eds.): From brows to trust: Evaluating embodied conversational agents. Kluwer Academic Publishers, Norwell, USA (2004)
45. Schank, R.C.: Memory based expert systems. AFOSR.TR. 84-0814, Comp. Science Dept., Yale University (1984)
46. Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Maceachren, A.M., Sengupta, K.: Speech-gesture driven multimodal interfaces for crisis management. Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1327-1354
47. Sharp, H., Rogers, Y., Preece, J.: Interaction Design, 2nd edition. John Wiley & Sons, Chichester, UK (2007)
48. Shneiderman, B.: Universal usability. Communications of the ACM, Vol. 43, No. 5 (2000) 85-91
49. Shneiderman, B.: CUU: Bridging the Digital Divide with Universal Usability. ACM Interactions, Vol. 8, No. 2 (2001) 11-15
50. Shoham, Y.: What we talk about when we talk about software agents. IEEE Intelligent Systems and Their Applications, Vol. 14, No. 2 (1999) 28-31
51. Tennenhouse, D.: Proactive computing. Communications of the ACM, Vol. 43, No. 5 (2000) 43-50
52. Turk, M.: Computer vision in the interface. Communications of the ACM, Vol. 47, No. 1 (2004) 61-67
53. Vo, M.T., Waibel, A.: Multimodal human-computer interaction. Proc. Int'l Symposium on Spoken Dialogue (1993)
54. Waibel, A., Vo, M.T., Duchnowski, P., Manke, S.: Multimodal Interfaces. Artificial Intelligence Review, Vol. 10, No. 3-4 (1995) 299-319
55. Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., Huang, T.S.: Audio-Visual Emotion Recognition in Adult Attachment Interview. Proc. Int'l Conf. Multimodal Interfaces (2006) 139-145
56. Zhang, P., Li, N.: The importance of affective quality. Communications of the ACM, Vol. 48, No. 9 (2005) 105-108

SmartWeb Handheld — Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services

Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, and Norbert Reithinger

German Research Center for Artificial Intelligence
66123 Saarbrücken, Germany
firstname.lastname@dfki.de

Abstract. SmartWeb aims to provide intuitive multimodal access to a rich selection of Web-based information services. We report on the current prototype with a smartphone client interface to the Semantic Web. An advanced ontology-based representation of facts and media structures serves as the central description for rich media content. Underlying content is accessed through conventional web service middleware to connect the ontological knowledge base and an intelligent web service composition module for external web services, which is able to translate between ordinary XML-based data structures and explicit semantic representations for user queries and system responses. The presentation module renders the media content and the results generated from the services and provides a detailed description of the content and its layout to the fusion module. The user is then able to employ multiple modalities, like speech and gestures, to interact with the presented multimedia material in a multimodal way.

1 Introduction

The development of a context-aware, multimodal mobile interface to the Semantic Web [1], i.e., ontologies and web services, is a very interesting task since it combines many state-of-the-art technologies such as ontology development, distributed dialog systems, standardized interface descriptions (EMMA[1], SSML[2], RDF[3], OWL-S[4], WSDL[5], SOAP[6], MPEG7[7]), and composition of web services. In this contribution we describe the intermediate steps in the dialog system development process for the project SmartWeb [2], which was started in 2004 by partners in industry and academia.

[1] http://www.w3.org/TR/emma
[2] http://www.w3.org/TR/speech-synthesis
[3] http://www.w3.org/TR/rdf-primer
[4] http://www.w3.org/Submission/OWL-S
[5] http://www.w3.org/TR/wsdl
[6] http://www.w3.org/TR/soap
[7] http://www.chiariglione.org/mpeg

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 272–295, 2007.

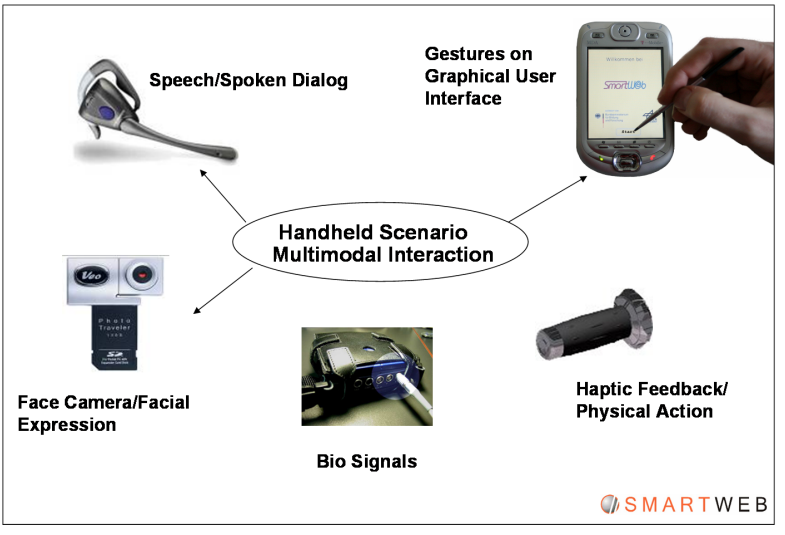


Fig. 1. The multimodal dialog handheld scenario comprises spoken dialog recorded by a bluetooth micro, gestures on the graphical PDA touchscreen, and camera signals. In addition, the SmartWeb project uses recognition of user state in biosignals to adapt system output in stressed car driving situations and haptic input from a force-feedback device installed on a motorbike.

In our main scenario, the user carries a smartphone PDA, as shown in figure 1, and poses closed and open domain multimodal questions in the context of football games and a visit to a Football Worldcup stadium. The PDA serves as an easy-to-use user interaction device which can be queried by natural language speech or handwriting, and which can understand social signaling - hand gestures on the PDA touchscreen and head movement perceived by the PDA camera. By our multimodal dialog interface we aim at providing natural interaction for human users in the Human Computing paradigm [3].

Many challenging tasks, such as interaction design for mobile devices with restricted computing power, have to be addressed: the user should be able to use the PDA as a question answering (QA) system, using speech and gestures to ask for information about players or games stored in ontologies, or other up-to-date information like weather forecast information accessible through web services, Semantic Web pages (Web pages wrapped by semantic agents), or the Internet.

The partners of the SmartWeb project share experience from earlier dialog system projects [4,5,6,7]. We followed guidelines for multimodal interaction, as explained in [8] for example, in the development process of our first demonstrator system [9] which contains the following assets: *multimodality*, more modalities allow for more natural communication, *encapsulation*, we encapsulate the multimodal dialog interface proper from the application, *standards*, adopting to standards opens the door to scalability, since we can re-use ours as well as other's resources, and *representation*. A shared representation and a common

ontological knowledge base eases the data flow among components and avoids costly transformation processes. In addition, semantic structures are our basis for representing dialog phenomena such as multimodal references and user queries. The same ontological query structures are input into the knowledge retrieval and web service composition process.

In the following chapters we demonstrate the strength of Semantic Web technology for information gathering dialog systems, especially the integation of multiple dialog components, and show how knowledge retrieval from ontologies and web services can be combined with advanced dialogical interaction, i.e., system-initiative callbacks, which present a strong advancement to traditional QA systems. Traditional QA realizes like a traditional NLP dialog system a (recognize) - analyze - react - generate - (synthesize) pipeline [10]. Once a query is started, the information is pipelined until the end, which means that the user-system interaction is reduced to user and result messages. The types of dialogical phenomena we address and support include reference resolution, system-initiated clarification requests and pointing gesture interpretation, among others. Support for underspecified questions and enumeration question types additionally shows advanced QA functionality in a multimodal setting. One of the main contributions is the ontology-based integration of verbal and non-verbal system input (fusion) and output (system reaction). System-initiative clarification requests and other pro-active or mixed-initiative system behaviour are representative for emerging multimodal and embedded HCI systems. Challenges for the evaluation of emerging Human Computing applications [11] traces back to challenges in multimodal dialog processing, such as error-prone perception and intergration of multimodal input channels [12,13,14]. Ontology-based integration of verbal and non-verbal system input and output can be seen as groundwork for robust processing of multimodal user input.

The paper is organized as follows: we begin with an example interaction sequence, in section 3, we explain the dialog system architecture. Section 4 describes the ontological knowledge representation, and section 5 the Web Service access. Section 6 then gives a description of the underlying ontology-based language parsing and discourse processing steps as well as their integration into a robust demonstrator system suitable for exhibitions such as CeBIT. Conclusions about the success of the system so far and future plans are outlined in section 7.

2 Multimodal Interaction Sequence Example

The following interaction sequence is typical for the SmartWeb dialog system.

---

(1) **U:** “When was Germany world champion?”
(2) **S:** “In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy), 2003 (in USA)”
(3) **U:** “And Brazil?”

(4) **S:** "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)" + [*team picture, MPEG-7 annotated*]
(5) **U:** Pointing gesture on player *Aldair* + "How many goals did this player score?"
(6) **S:** "Aldair scored none in the championship 2002."

(7) **U:** "What can I do in my spare time on Saturday?"
(8) **S:** "Where?"
(9) **U:** "In Berlin."
(10) **S:** *The cinema program, festivals, and concerts in Berlin are listed.*

The first and second enumeration questions are answered by deductive reasoning within the ontological knowledge base modeled in OWL [15] representing the static but very rich implicit knowledge that can be retrieved. The second example beginning with (7) evokes a dynamically composed web service lookup. It is important to note that the query representation is the same for all the access methods to the Semantic Web (cf. section 6.1) and is defined by foundational and domain-specific ontologies. In a case where the GPS co-ordinates were accessible from the mobile device, the clarification question would have been omitted.

3 Architecture Approach

A flexible dialog system platform is required in order to allow for true multi-session operations with multiple concurrent users of the server-side system as well as to support audio transfer and other data connections between the mobile device and a remote dialog server. These types of systems have been developed, like the Galaxy Communicator [16] (cf. also [17,18,19,20]), and commercial platforms from major vendors like VoiceGenie, Kirusa, IBM, and Microsoft use X+V1, HTML+SALT2, or derivatives for speech-based interaction on mobile devices. For our purposes these platforms are too limited. To implement new interaction metaphors and to use Semantic Web based data structures for both dialog system internal and external communication, we developed a platform designed for Semantic Web data structures for NLP components and backend knowledge server communication. The basic architecture is shown in figure 2.

It consists of three basic processing blocks: the PDA client, the dialog server, which comprises the dialog manager, and the Semantic Web access system.

On the PDA client, a local Java-based control unit takes care of all I/O, and is connected to the GUI-controller. The local VoiceXML-based dialog system resists on the PDA for interaction during link downtimes.

The dialog server system platform instantiates one dialog server for each call and connects the multimodal recognizer for speech and gesture recognition. The

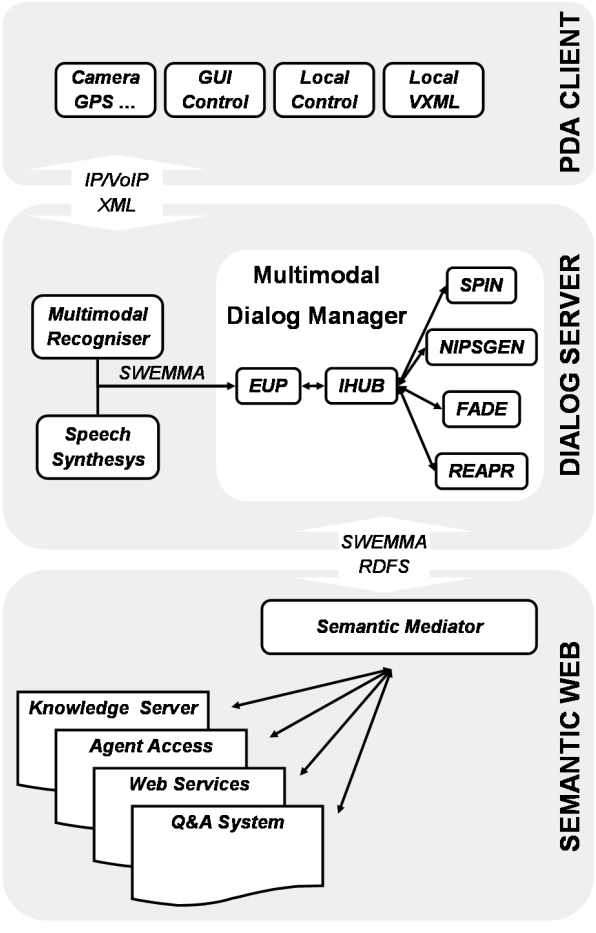


Fig. 2. SmartWeb handheld architecture

dialog system instantiates and sends the requests to the *Semantic Mediator*, which provides the umbrella for all different access methods to the Semantic Web we use. It consists of an open domain QA system, a Semantic Web service composer, Semantic Web pages (wrapped by semantic agents), and a knowledge server.

The dialog system consist of different, self-contained processing components. To integrate them we developed a Java-based hub-and-spoke architecture [21]. The most important processing modules in the dialog system connected to the IHUB are: a speech interpretation component (SPIN), a modality fusion and discourse component (FADE), a system reaction and presentation component (REAPR), and a natural language generation module (NIPSGEN), all discussed in section 6. An EMMA Unpacker/Packer (EUP) component provides the communication with the dialog server and Semantic Web subsystem external to the

multimodal dialog manager and communicates with the other modules of the dialog server, the multimodal recognizer, and the speech synthesis system.

Processing a user turn, normal data flows through $SPIN \rightarrow FADE \rightarrow REAPR \rightarrow SemanticMediator \rightarrow REAPR \rightarrow NIPSGEN$. However, the data flow is often more complicated when, for example, misinterpretations and clarifications are involved.

4 Ontology Representation

The ontological infrastructure of the SmartWeb dialog system project, the SWIntO (**S**mart**W**eb **Int**egrated **O**ntology) [22], is based on an upper model ontology realized by merging well chosen concepts from two established foundational ontologies, DOLCE [23] and SUMO [24], into a unique one: the SmartWeb foundational ontology SmartSUMO [25]. Domain specific knowledge (sportevent, navigation) is defined in dedicated ontologies modeled as subontologies of the SmartSUMO. The SWIntO integrates question answering specific knowledge of a discourse ontology (DiscOnto) and representation of

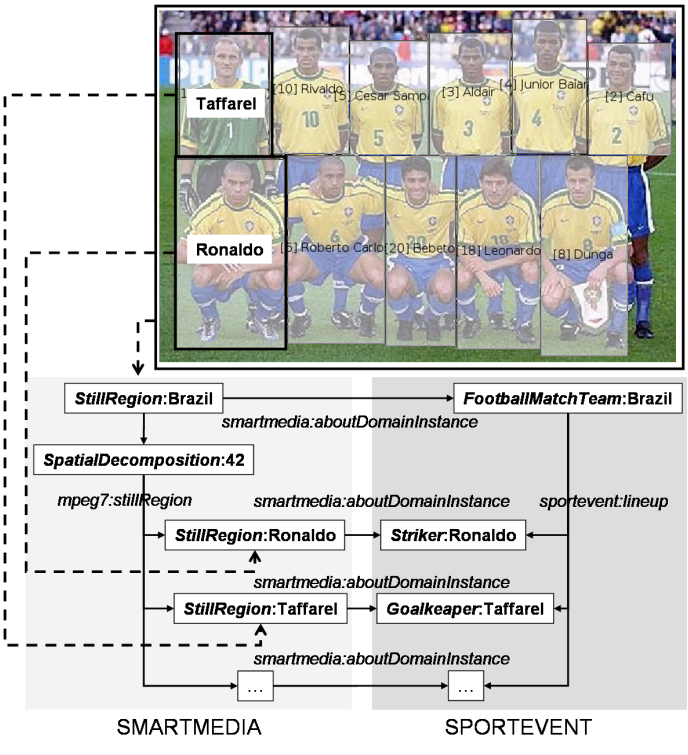


Fig. 3. A SmartMedia instance representing the decomposition of the Brazil 1998 world cup football team image

multimodal information of a media ontology (SMARTMEDIA). The data exchange is RDF-based.

4.1 The Upper Model

In order to integrate knowledge from different domains we created an abstract foundational ontology that models basic properties and a common relational background for interoperability. Each domain specific ontology has been then aligned to this foundational ontology. The process of creation of SMARTSUMO has been constrained by two essential principles: the ontology must offer a *rich axiomatization* with a high abstraction level and cover a large number of general concepts. The ontology should also be *descriptive* for modeling artifacts of human common sense and give the possibility of modeling entities extended in time and space. Therefore the SMARTSUMO requires a *perdurantism* approach. From about a dozen freely available foundational ontologies (see [26] for an overview) the DOLCE and the SUMO ontology were selected as being the best fit for the given task. Both ontologies have been modified and combined. The upper level of SMARTSUMO is basically derived from DOLCE with the distinction between *endurant*, *perdurant*, *abstract* and *qualities*, and the rich axiomatisation that allows the modelling of location in space and time, and the use of relations such as parthood and dependence. We also borrowed the ontology module *Descriptions & Situations* [27] from DOLCE. From this minimal core of generic concepts we aligned the rich SUMO taxonomy.

4.2 The DiscOnto Ontology

We created a discourse ontology (DISCONTO) with particular attention to the modeling of discourse interactions in QA scenarios. The DISCONTO provides concepts for dialogical interaction with the user as well as more technical request-response concepts for data exchange with the Semantic Web subsystem including answer status, which is important in interactive systems. In particular DISCONTO comprises concepts for multimodal dialog management, a dialog act taxonomy, lexical rules for syntactic-semantic mapping, HCI concepts (e.g. pattern language for interaction design [28]), and concepts for questions, question focus, semantic answer types [29], and multimodal results [30].

Information exchange between the components of the server-side dialog system is based on the W3C EMMA standard that is used to realize containers for the ontological instances representing, e.g., multimodal input interpretations. SWEMMA is our extension of the EMMA standard which introduces additional *Result* structures in order to represent components output. On the ontological level we modeled an RDF/S-representation of EMMA/SWEMMA.

4.3 The Smartmedia Ontology

The SMARTMEDIA is an MPEG7-based media ontology and an extension to [31,32] that we use to represent output result, offering functionality for multimedia decomposition in space, time and frequency (mpeg7:SegmentDecomposition),

file format and coding parameters (mpeg7:MediaFormat), and a link to the Upper Model Ontology (smartmedia:aboutDomainInstance). In order to close the semantic gap between the different levels of media representations, the *smartmedia:aboutDomainInstance* property has been located in the top level class *smartmedia:Segment*. The link to the upper model ontology is inherited to all segments of a media instance decomposition to guarantee deep semantic representations for the *smartmedia* instances referencing the specific media object and for making up segment decompositions [33].

Figure 3 shows an example of this procedure applied to an image of the Brazilian football team in the final match of the World Cup 1998, as introduced in the interaction example. In the example an instance of the class *mpeg7:StillRegion*, representing the complete image, is decomposed into different *mpeg7:StillRegion* instances representing the segments of the image which show individual players.

The *mpeg7:StillRegion* instance representing the entire picture is then linked to a *sportevent:MatchTeam* instance, and each segment of the picture is linked to a *sportevent:FieldFootballPlayer* instance or sub-instance. These representations offer a framework for gesture and speech fusion when users interact with Semantic Web results such as MPEG7-annotated images, maps with points-of interest, or other interactive graphical media obtained from the ontological knowledge base or multimedia web services.

5 Multimodal Access to Web Services

To connect to web services we developed a semantic representation formalism based on OWL-S and a service composition component able to interpret an ontological user query. We extended the OWL-S ontologies to flexibly compose and invoke web services on the fly, gaining sophisticated representation of information gathering services fundamental to SmartWeb.

Sophisticated data representation is the key for developing a composition engine that exploits the semantics of web service annotation and query representation. The composition engine follows a plan-based approach as explained, e.g., in [34]. It infers the initial and goal state from the semantic representation of the user query, whereas the set of semantic web services is considered as planning operators. The output gained from automatic web service invocation is represented in terms of instances of the SmartWeb domain ontologies and enriched by additional media instances, if available. Media objects are represented in terms of the Smartmedia ontology (see above) and are annotated automatically during service execution. This enables the dialog manager for multimodal interaction with web service results.

A key feature of the service composition engine is to detect underspecified user queries, i.e., the lack of required web service input parameters. In these cases the composition engine is able to formulate a clarification request as specified within the discourse ontology (DiscOnto). This points out the missing pieces of

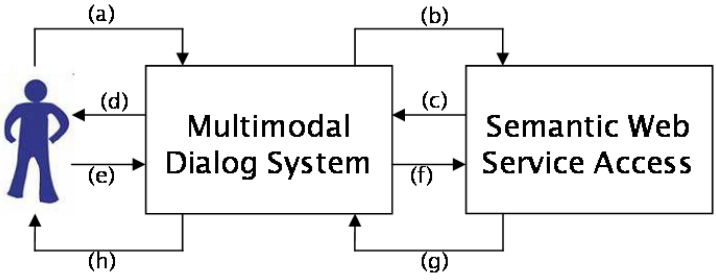


Fig. 4. Data flow for the processing of a clarification request as in the example (7-10) "What can I do in my spare time on Saturday?"

information to be forwarded to the dialog manager. Then the composition engine expects a clarification response enabling it to replan the refined ontological user query.

According to the interaction example (7-10) the composition engine searches for a web service demanding an activity event type and gets its description. Normally, the context module incorporated in the dialog manager would complete the query with the venue obtained from a GPS receiver attached to the handheld device. In the case of no GPS signal, for instance indoors, the composition engine asks for the missing parameter (cf. figure 4), which makes the composition engine more robust and thus more suitable for interactive scenarios.

In the interaction example (7-10) the composition planner considers the *T-Info EventService* appropriate for answering the query. This service requires both date and location for looking up events. While the date is already mentioned in the initial user query, the location is then asked of the user through clarification request. After the location information (dialog step (9) in the example:*In Berlin*) is obtained from the user, the composition engine invokes in turn two T-Info (DTAG) web services[8] offered by Deutsche Telekom AG (see also [35]): first the *T-Info EventService* as already mentioned above, and then the *T-Info MapService* for calculating an interactive map showing the venue as point-of-interest. Text-based event details, additional image material, and the location map are semantically represented (the map in MPEG7) and returned to the dialog engine.

[8] http://services.t-info.de/soap.index.jsp

6 Semantic Parsing and Discourse Processing

Semantic parsing and other discourse processing steps are reflected on the interaction device as advanced user perceptual feedback functionality. The following screenshot illustrates the two most important processing steps for system-user interaction, the feedback on the natural language understanding step and the presentation of multimodal results. The semantic parser produces a semantic query (illustrated on the left in figure 5), which is presented to the user in nested attribute-value form. The web service results (illustrated on the right in figure 5) for the interaction example (7-10) are presented in a multimodal way, combining text, image, and speech: *5 Veranstaltungen* (five events).

Fig. 5. Semantic query (illustrated on the left) and web service results (illustrated on the right)

6.1 Language Understanding with SPIN and Text Generation with NIPSGEN

Language Understanding

The parsing module is based on the semantic parser SPIN [36]. A syntactic analysis of the input utterance is not performed, but the ontology instances are created directly from word level. The typical advantages of a semantic parsing approach are that processing is faster and more robust against speech recognition errors and disfluencies produced by the user and the rules are easier to write and maintain. Also, multilingual dialog systems are easier to realize, as a syntactic analysis is not required for each supported language. A disadvantage is that

the complexity of the possible utterances is somewhat limited, but – in our experience – this is acceptable for dialog systems.

Several semantic parsers were developed for spoken dialog systems. Most of them use as underlying formalisms context free grammars (CFGs), e.g., [37] or finite state transducers (FSTs), e.g., [38] or variants of them, e.g., [39,40].

The SPIN parser uses a more powerful rule language to simplify writing of rules and to reduce the amount of required rules.

Properties of the rule language include:

- Direct handling of nested typed feature structure is available, which is important for processing more complex utterances.
- Order-independent matching is supported, i.e., the order of matched input elements is not important. This feature helps with the processing of utterances in free word order languages, like German, Turkish, Japanese, Russian or Hindi, and simplifies the writing of rules that are robust against speech recognition errors and disfluencies produced by the user. The increased robustness is achieved, as the parts of the utterance that are recognized incorrectly can be skipped. This is a mechanism that is also used in other approaches, e.g., [41].
- Built-in support for referring expressions is available.
- Regular expressions are available. Formulating the rules in a more elegant way is supported by this feature whereby the amount of required rules is reduced. Furthermore, the writing of robust rules is simplified.
- Constraints over variables and action functions are supported providing enough flexibility for real-world dialog system. Especially, if the ontology is developed without the parsing module in mind, flexibility is highly demanded.

SPIN's powerful rule language requires an optimizing parser, otherwise processing times would not be acceptable. Principally, the power of the rule language avoids the development of a parser which delivers sufficient performance for an arbitrary rule set. In particular, order-independent matching makes efficient parsing much harder, parsing of arbitrary grammars is NP-complete, see also [42]. Therefore, the parser is tuned for rule sets that are typical for dialog systems. A key feature achieving fast processing is the pruning of results that can be regarded as irrelevant for further processing within the dialog system. Pruning of results means that the parsing algorithm is not complete. Pruning of irrelevant results is achieved using a fixed application order for the rules in combination with tagging some of the rules as destructive. More details of the parsing algorithm can be found in [36]. The rule set used for the SmartWeb project consists of 1069 rules where 363 rules are created manually, and 706 are generated automatically from the linguistic information stored in SWIntO, e.g., country names. The lexicon contains 2250 entries. Currently, the knowledge base for the SmartWeb system consists of 1069 rules whereby 363 rules were created manually, and 706 were generated automatically from the linguistic information stored in SWIntO, e.g., country names. The lexicon contains 2250 entries. The average processing time is about 50ms per utterance, which ensures direct feedback to user inputs.

To demonstrate processing of rules, four rules are provided as examples of how to process the utterance *When was Brazil world champion?*. The first one transforms the word *Brazil* into the ontology instance `Country`:

```
Brazil
→ Country(name:BRAZIL)
```

The second rule transforms countries to teams, as each country can stand for a team in our domain:

```
$C=Country()
→ Team(origin:$C)
```

The third rule processes *when*, generating an instance of the type `TimePoint` which is marked as questioned:

```
when
→ TimePoint(variable:QEVariable(focus:text))
```

The fourth rule processes the verbal phrase *<TimePoint> was <Team> world champion*

```
$TP=TimePoint() was $TM=Team() world champion
→ QEPattern(patternArg:Tournament(
                winner:$TM, happensAt:$TP))
```

Text Generation

Within the dialog system, the text generation module is used within two processing steps. First, the abstract interpretation of the user utterance is shown as human readable text, called paraphrase. This allows the user to check if the query has been interpreted with the desired meaning and if all of the provided information has been included. Second, the search results represented as instances of SWIntO are verbalized.

The text generation module uses the same SPIN parser that is used in the language understanding module together with a TAG (tree adjoining grammar) grammar module [43]. The TAG grammar in this module is derived from the XTAG grammar for English developed at the University of Pennsylvania.[9]

The inputs of the generation module are instances of SWIntO representing the search results. Then these results are verbalized in different ways, e.g., as a heading, as an image description, as a row of a table, or as a text which is synthesized. A processing option indicates the current purpose. Figure 6 shows an example containing different texts.

The input is transformed into an utterance in four steps:

1. An intermediate representation is built up on a phrase level. The intermediate representation is introduced, as a direct generation of the TAG tree description would lead to overly complicated and unintuitive rules. The required rules are domain dependent.

[9] http://www.cis.upenn.edu/∼xtag/

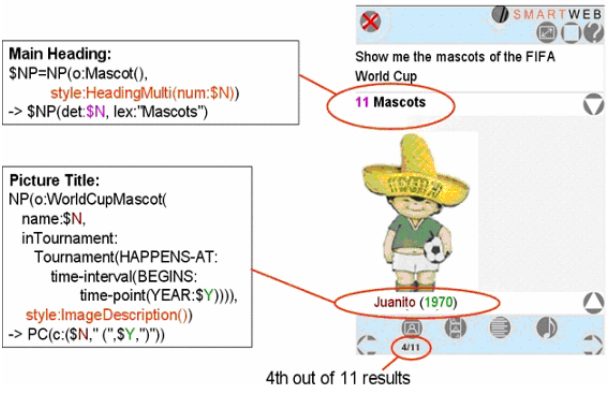


Fig. 6. The verbalized result for the utterance *Show me the mascots of the FIFA World Cup*. The rules generating the main heading and the image description are shown on the left.

2. A set of domain independent rules transforms the intermediate representation to a derivation tree for the TAG-grammar. Each tree in the TAG grammar has a corresponding type in the ontology of the text generation module. The features of a TAG tree type represent the type of operation (adjunction (`a`), substitution (`s`), lexical replacement (`l`)) and the position in the tree, e.g., `211`.
3. The actual syntax tree is constructed using the derivation tree. After the tree has been built up, the features of the tree nodes are unified.
4. The correct inflections for all lexical leafs are looked up in the lexicon. Traversing the lexical leafs from left to right produces the result text.

For text generation, the parser is driven in a slightly different mode: The automatic ordering of rules is switched off, instead the order in which the rules are applied is taken from the file containing the rules. Regions that have to be applied in a loop and rules that have to be applied optionally are marked explicitly. In the current system, two loops exists, one for each phase. A more detailed description of the text generation module can be found in [44].

In the SmartWeb system currently 179 domain dependent generation rules and 38 domain independent rules are used.

6.2 Multimodal Discourse Processing with FADE

An important aspect of SmartWeb is its context-aware processing strategy. All recognized user actions are processed with respect to their situational and discourse context. A user is thus not required to pose separate and unconnected questions. In fact, she might refer directly to the situation, e.g., *"How do I get to Berlin from here?"*, where *here* is resolved from GPS information, or to previous contributions (as in the elliptical expression *"And in 2002?"* in the context of a previously posed question *"Who won the Fifa World Cup in 1990?"*). The

interpretation of user contributions with respect to their discourse context is performed by a component called *Fusion and Discourse Engine*—FADE [45,46][10]. The task of FADE is to integrate the verbal and nonverbal user contributions into a coherent multimodal representation to be enriched by contextual information, e.g., resolution of referring and elliptical expressions.

The basic architecture of FADE consists of two interwoven processing layers: (1) a production rule system—PATE—that is responsible for the reactive interpretation of perceived monomodal events, and (2) a discourse modeler—DiM—that is responsible for maintaining a coherent representation of the ongoing discourse and for the resolution of referring and elliptical expressions.

In the following two subsections we will briefly discuss some context-related phenomena that can be resolved by FADE.

Resolution of referring expressions. A key feature of the SmartWeb system is that the system is capable of dealing with a broad range of referring expressions as they occur in natural dialogs. This means the user can employ deictic references that are accompanied by a pointing gesture (such as in *"How often did this team [pointing gesture] win the World Cup?"*) but also—if the context provides enough disambiguating information—without any accompanying gestures (e.g., if the previous question is uttered in the context of a previous request like *"When was Germany World Cup champion for the last time?"*).

Moreover, the user is also able to utter time deictic references as in *"What's the weather going to be like tomorrow?"* or *"What's the weather going to be like next Saturday?"*.

Another feature supported by FADE is the resolution of *cross modal* spatial references, i.e., a spoken reference to visually displayed information. The user can refer, for example, to an object that is currently displayed on the screen. If a picture of the German football team is displayed, the system is able to resolve references like *"this team"* even when the team has not yet been mentioned verbally. MPEG7-annotated images (see section 4) even permit spatial references to objects displayed within pictures, e.g., as in *"What's the name of the guy to the right of Ronaldo?"* or *"What's the name of the third player in the top row?"*.

Resolution of elliptical expression. Humans tend to keep their contributions as short and efficient as possible. This is particularly the case for follow-up questions or answers to questions. Here, people often make use of elliptical expressions, e.g., when they ask a follow-up question *"And the day after tomorrow?"* in the context of a previous question *"What's the weather going to be like tomorrow?"*. But even for normal question-answer pairs people tend to omit everything that has already been conveyed by the question (User: *"Berlin"* in the context of a clarification question of the system like *"Where do you want to start?"*; see section 5).

Elliptical expressions are processed in SmartWeb as follows: First, SPIN generates an ontological query that contains a semantic representation of the

[10] The situational context is maintained by another component called *SitCom* that is not discussed in this paper (see [47]).

elliptical expression, e.g., in case of the aforementioned example “Berlin”. This analysis would only comprise an ontological instance representing the city Berlin. FADE in turn, then tries to integrate the elliptical expression with the previous system utterance, if this was a question. Otherwise it tries to integrate the elliptical expression with the previous user request. If the resolution succeeded, the resulting interpretation either describes the answer to the previous clarification question, or it describes a new question.

OnFocus/OffFocus identification. An important task for mobile, speech-driven interfaces that support an open-microphone[11] is the continuous monitoring of all input modalities in order to detect when the user is addressing the system. In the mobile scenario of SmartWeb, the built-in camera of the MDA Pro handheld can be used to track whether a user is present. This camera constantly captures pictures of the space immediately infront of the system. These pictures are processed by a server-side component that detects whether the user is looking at the device or not.

In SmartWeb, there are two components that determine the attentional state of the user: (i) the OnView recognizer, and the (ii) the OnTalk recognizer. The task of the OnView recognizer is to determine whether the user is looking at the system or not. The OnView-Recognizer analyzes a video signal captured by a video camera linked to the mobile device and determines for each frame whether the user is in OnView or OffView mode (figure 7 shows two still images of these different modes).

Fig. 7. Two still images illustrating the function of the OnView/OffView recognizer: The image on the left shows the OnView case and the one the right shows the Offview case

The task of the OnTalk recognizer is to determine whether a user’s utterance is directed to the system. To this end, the OnTalk recognizer analyzes the speech signal and computes about 99 prosodic features based on F0, energy, duration,

[11] Open-microphone means the microphone is always active so that the user can interact with the system without further activation. In contrast to an open-microphone interface, systems often require the user to push some hard- or software button in order to activate the system (i. e., a *push-to-activate* button).

jitter, and shimmer (see [48]). This is done for each word but the final result is averaged over the complete turn. Both recognizers provide a score reflecting the individual confidence of a classification.

FADE receives and stores the results of the OnView recognizers as a continuous stream of messages (i. e., every time the OnView state changes, FADE receives an update). OnTalk/OffTalk classifications are only sent to FADE if the speech recognition components detected some input event. The actual algorithm goes as follows: The overall idea is to combine the two distinct classifications for OnView/OffView and OnTalk/OffTalk in order to compensate for potential classification errors. If the OnView value is above 0.3 (where 0 means OffView and 1 means OnView), the OffTalk value must be very low (below 0.2) in order to classify a contribution as OffFocus. Otherwise, a OnTalk value below 0.5 is already sufficient to classify an utterance as OffFocus.

6.3 Reaction and Presentation Planning for the Semantic Web

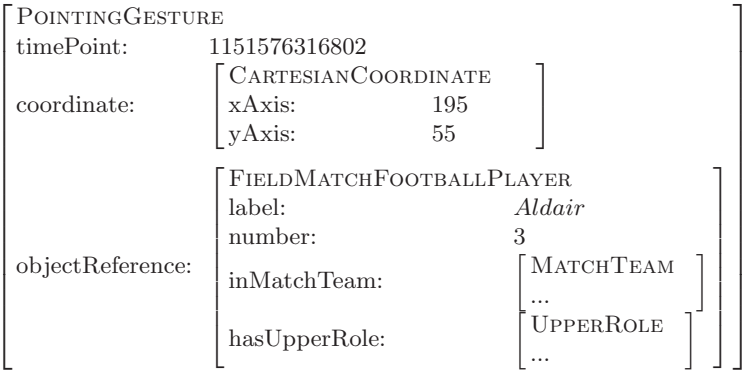
An integral part of dialog management is the reaction and presentation module (REAPR). It manages the dialogical interaction for the supported dialog phenomena such as flexible turn-taking, incremental processing, and multimodal fusion of system output. REAPR is based on a finite-state-automaton and information space (IS). The FSA makes up the integral part of the dialog managment decisions in the specific QA domain we model. The dialog structure that is embedded and committed by the transitions of the FSA allows for a declarative control mechanism for reaction and presentation behaviour.

Our new approach differs from other IS approaches (e.g. [49]) by generating IS features from the ontological instances generated during dialog processing [50].[12]

Since the dialog ontology is a model for multimodal interaction, multimodal MPEG7 result representations, multimodal result presentations, dialog state, and (agent) communication with the backend knowlege servers, large information spaces can be extracted from the ontological instances describing the system and user turns in terms of special dialog acts - to ensure accurate dialog management capabilities. REAPR decides, for example, if a semantic query is acceptable for transfer to the Semantic Mediator. The IS approach to dialog modeling comprises, apart from dialog moves and update strategies, a description of informational components (e.g. common ground) and their formal representations. Since in REAPR the formal dialog specification consists of ontological Semantic Web data structures, a formal well-defined complement to previous formal logic-based operators and Discourse Representation Structures (DRS) is provided. However, the ontological structures resemble the typed feature structures (TFS) [51] that we use for illustration further down. During interaction, many message transfer processes take place, mainly for query recognition and query processing, all of which are based on Semantic Web ontological structures, and REAPR is involved

[12] The IS state is traditionally divided into global and local variables which make up the knowledge state at a given time point. Ontological structures that change over time vastly enhance the representation capabilities of dialog management structures, or other structures like queries from which relevant features can also be extracted.

in many of them. Here we give an example of ontological representations of user pointing gestures (dialog step (5) in the interaction example) which are obtained from the PDA and transformed into ontology-structures to be used by the input fusion module. The following figure shows the ontological representation of a pointing gesture as TFS.



It is important to mention that dialog reaction behaviour within SmartWeb is governed by the general QA scenario, which means that almost all dialog and system moves relate to questions, follow-up questions, clarifications, or answers. As these dialog moves can be regarded as adjacency pairs, the dialog behaves according to some finite state grammar for QA, which makes up the automaton part (FSA) in REAPR. The finite state approach enhances robustness and portability and allows to demonstrate dialog management capabilities even before the more complex IS states are available to be integrated into the reaction and presentation decision process.

6.4 Information States for QA

Information state theory of dialog modelling consists basically of a description of informal components (e.g., obligations, beliefs, desires, intentions) and their formal representation [52]. IS states as envisioned here do not declare update rules and an update strategy (for e.g. discourse obligations [53]) because the data-driven approach is pattern-based, using directly observable processing features, which complements an explicit manual formulation of update rules. Since the dialog ontology is a formal representation model for multimodal interaction, multimodal MPEG-7 result representations [30], result presentations [28], dialog state, and (agent) communication with the backend knowledge servers, large information spaces can be extracted from the ontological instances describing the system and user turns in terms of realized dialog acts.

The turn number represents our first FSA extension to IS with the result of increased flexibility to user replies. Replies which are not specified in a pathway,

Table 1. IS Feature Classes and Features

Feature Class	IS State Features
MMR	*Listening, Recording, Barge-in, Last-ok, Input dominance (text or voice)*
NLU	*Confidence, Domain relevance*
Query	*Dialog act, Focus medium, Complexity, Context object, Query text*
Fusion	*Fusion act, Co-reference resolution*
Answer	*Success, Speed, Answer streams, Status, Anser type, Content, Answer text*
Manager	*Turn/Task numbers, Idle states, Waiting for Results, User/system turn, Elapsed times: input/output, Dialog act history (system and user) e.g. reject, accept, clarify*

are not considered erroneous by default, since the IS now contains a new turn value. Ontologicial features for IS extraction under investigation are summarised in table 1.

In previous work on dialog management adaptations [54,55,56], reinforcement learning was used, but large state spaces with more than about five non-binary features are still hard to deal with. As seen in table 1, more than five relevant features can easily be declared. Since our optimisation problem can be formulated at very specific decisions in dialog management due to the FSA ground control, less training material for larger feature extractions is to be expected.

Relevance selection of ontology-based features is the next step for ontology-based dialog management adaptations. In the context of Human Computing one question is how prior user knowledge can be incorporated in order to select relevant features so as to converge faster toward more effective and natural dialog managers. We already incorporated human dialog knowledge by the dialog FSA structure. In a more user-centered and dynamic scenario, the user in the loop should accelerate Learning [57] by e.g. selecting the IS features that are most relevant in the specific dialog application. The human-user interaction for this selection process is of particular interest in dialog applications. Is it possible to integrate the feature selection process into a normal dialog session that the user and the dialog system engage in? In the context of the SmartWeb project we will develop a tool to run the dialog system with the additional possibility to interfere in the dialog management in case the user is not satisfied with the processing. Our future plans include measuring when the direct user feedback is likely to be useful for adapting dialog management strategies automatically. One example is to generate useful reactions in cases where the natural language understanding component fails. Whenever there is the freedom to formulate statements, which is a precondition for natural language communication, understanding may be difficult. What can be done in such cases is to produce useful

reactions and to give hints to the user or examples that the use of supported terminology is not insisted, but at least directed.

6.5 Dialog Components Integration

In this section we will focus on issues of interest pertaining to the system integration. In the first instance, dialog component integration is an integration on a conceptual level. All dialog manager components communicate via ontology instances. This assumes the representation of all relevant concepts in the foundational and domain ontologies – which is hard to provide at the beginning of the integration. In our experience, using ontologies in information gathering dialog systems for knowledge retrieval from ontologies and web services in combination with advanced dialogical interaction is an iterative ontology engineering process. This process requires very disciplined ontology updates, since changes and extensions must be incorporated into all relevant components. The additional modeling effort pays off when regarding the strength of this Semantic Web technology for larger scale projects.

We first built up an initial discourse ontology of request-response concepts for data exchange with the Semantic Web sub-system. In addition, an ontological dialog act taxonomy has been specified, to be used by the semantic parsing and discourse processing modules. A great challenge is the mapping between semantic queries and the ontology instances in the knowledge base. In our system, the discourse (understanding) specific concepts have been linked to the foundational ontology and, e.g., the sportevent ontology, and the semantic parser only builds up interpretations with SWIntO concepts. Although this limits the space of possible interpretations according to the expressivity of the foundational and domain ontologies, the robustness of the system is increased. We completely circumvent the problem of concept and relation similarity matching between conventional syntactic/semantic parsers and backend retrieval systems.

Regarding web services we transform the output from the web services, in particular maps with points of interest, into instances of the SmartWeb domain ontologies for the same reasons of semantic integration. As already noted, ontological representations offer a framework for gesture and speech fusion when users interact with Semantic Web results such as MPEG7-annotated images and maps. Challenges in multimodal fusion and reaction planning can be addressed by using more structured representations of the displayed content, especially for pointing gestures, which contain references to player instances after integration. We extended this to pointing gesture representations on multiple levels in the course of development, to include representations of the interaction context, the modalities and display patterns used, and so on.

The primary aim is to generate structured input spaces for more context-relevant reaction planning to ensure naturalness in system-user interactions to a large degree. Currently, as shown in chapter 6.2, we are experimenting with the MDA's camera input indicating whether the user is looking at the device, to combine it with other indicators to a measure of user focus. The challenge of integrating and fusing multiple input modalities can be reduced by ontological

representations, which exist at well-defined time-points, and are also accessible to other components such as the semantic parser, or the reaction and presentation module.

7 Conclusions

We presented a mobile system for multimodal interaction with an ontological knowledge base and web services in a dialog-based QA scenario. The interface and content representations are based on W3C standards such as EMMA and RDF. The world knowledge shared in all knowledge-intensive components is based on the existing ontologies SUMO and DOLCE, for which we added additional concepts for QA and multimodal interaction in a discourse ontology branch.

We presented the development of the second demonstrator of the SmartWeb system which was successfully demonstrated in the context of the Football World Cup 2006 in Germany. The SWIntO ontology now comprises 2308 concept classes, 1036 slots and 90522 instances.[13] For inference and retrieval the ontology constitutes 78385 data instances after deductions.[14] The answer times are in a 1 to 15 seconds time frame for about 90% of all questions. In general, questions without images and videos as answers can be processed much faster. The web service composer addresses 25 external services from traveling (navigation, train connections, maps, hotels), event information, points of interest (POIs), product information (books, movies), webcam images, and weather information.

The SmartWeb architecture supports advanced QA functionalities such as flexible control flow to allow for clarification questions of web services when needed, long- and short-term memory provided by distributed dialog management in the fusion and discourse module and in the reaction and presentation module, as well as semantic interpretations provided by the speech interpretation module. This can be naturally combined with dialog system strategies for error recoveries, clarifications with the user, and multimodal interactions. Support for inferential, i.e., deductive reasoning, which we provide, complements the requirements for advanced QA in terms of information- and knowledge retrieval. Integrated approaches as presented here rely on ontological structures and a deeper understanding of questions, not at least to provide a foundation for result provenance explanation and justification. Our future plans on the final six month agenda include dialog management adaptations via machine learning and collaborative filtering of redundant results in our multi-user enviroment, and incremental presentation of results.

Acknowledgments

The research presented here is sponsored by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A (SmartWeb). We thank our

[13] The SWIntO can be downloaded at the SmartWeb homepage for research purposes.
[14] The original data instance set was 175293 instances, but evoked processing times up to two minutes for single questions by what interactivity was no longer guaranteed.

student assistants and the project partners. The responsibility for this papers lies with the authors.

References

1. Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W., eds.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. In Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W., eds.: Spinning the Semantic Web, MIT Press (2003)
2. Wahlster, W.: SmartWeb: Mobile Applications of the Semantic Web. In Dadam, P., Reichert, M., eds.: GI Jahrestagung 2004, Springer (2004) 26–27
3. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: a survey. In: ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces, New York, NY, USA, ACM Press (2006) 239–248
4. Wahlster, W., ed.: VERBMOBIL: Foundations of Speech-to-Speech Translation. Springer (2000)
5. Wahlster, W.: SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In Krahl, R., Günther, D., eds.: Proc. of the Human Computer Interaction Status Conference 2003, Berlin, Germany, DLR (2003) 47–62
6. Reithinger, N., Fedeler, D., Kumar, A., Lauer, C., Pecourt, E., Romary, L.: MIAMM - A Multimodal Dialogue System Using Haptics. In van Kuppevelt, J., Dybkjaer, L., Bernsen, N.O., eds.: Advances in Natural Multimodal Dialogue Systems. Springer (2005)
7. Wahlster, W.: SmartKom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
8. Oviatt, S.: Ten myths of multimodal interaction. Communications of the ACM **42**(11) (1999) 74–81
9. Reithinger, N., Bergweiler, S., Engel, R., Herzog, G., Pfleger, N., Romanelli, M., Sonntag, D.: A Look Under the Hood Design and Development of the First SmartWeb System Demonstrator. In: Proceedings of 7th International Conference on Multimodal Interfaces (ICMI 2005), Trento, Italy (October 04-06 2005)
10. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A.: An Architecture for a Generic Dialogue Shell. Natural Language Engineering **6**(3) (2000) 1–16
11. Poppe, R., Rienks, R.: Evaluating the future of hci: Challenges for the evaluation of upcoming applications. In: Proceedings of the International Workshop on Artificial Intelligence for Human Computing at the International Joint Conference on Artificial Intelligence IJCAI'07, Hyderabad, India (2007) 89–96
12. Oviatt, S.: Multimodal Interfaces. In: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Assoc. (2003) 286–304
13. Wasinger, R., Wahlster, W.: The Anthropomorphized Product Shelf: Symmetric Multimodal Interaction with Instrumented Environments. In Aarts, E., Encarnao, J.L., eds.: Chapter in: True Visions: The Emergence of Ambient Intelligence. Springer-Verlag, Berlin, Heidelberg, Germany (2006)

14. Wahlster, W.: Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In: KI. (2003) 1–18
15. Krotzsch, M., Hitzler, P., Vrandecic, D., Sintek, M.: How to reason with OWL in a logic programming system. In: Proceedings of RuleML'06. (2006)
16. Cheyer, A.J., Martin, D.L.: The Open Agent Architecture. Autonomous Agents and Multi-Agent Systems **4**(1–2) (2001) 143–148
17. Seneff, S., Lau, R., Polifroni, J.: Organization, Communication, and Control in the Galaxy-II Conversational System. In: Proc. of Eurospeech'99, Budapest, Hungary (1999) 1271–1274
18. Thorisson, K.R., Pennock, C., List, T., DiPirro, J.: Artificial intelligence in computer graphics: A constructionist approach. Computer Graphics (February 2004) 26–30
19. Herzog, G., Ndiaye, A., Merten, S., Kirchmann, H., Becker, T., Poller, P.: Large-scale Software Integration for Spoken Language and Multimodal Dialog Systems. Natural Language Engineering **10** (2004) Special issue on Software Architecture for Language Engineering.
20. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering **10** (2004) Special issue on Software Architecture for Language Engineering.
21. Reithinger, N., Sonntag, D.: An integration framework for a mobile multimodal dialogue system accessing the semantic web. In: Proc. of Interspeech'05, Lisbon, Portugal (2005)
22. Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Vembu, S., Baumann, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N., Loos, B., Porzel, R., Zorn, H.P., Micelli, V., Schmidt, C., Weiten, M., Burkhardt, F., Zhou, J.: Dolce ergo sumo: On foundational and domain models in swinto (smartweb integrated ontology). Technical report, AIFB, Karlsruhe (July 2006)
23. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schcneider, L.: Sweetening Ontologies with DOLCE. In: In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02). Volume 2473 of Lecture Notes in Computer Science., Sigünza, Spain (Oct. 1–4 2002) 166 ff
24. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In Welty, C., Smith, B., eds.: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine (October 17–19 2001)
25. Cimiano, P., Eberhart, A., Hitzler, P., Oberle, D., Staab, S., Studer, R.: The smartweb foundational ontology. Technical report, (AIFB), University of Karlsruhe, Karlsruhe, Germany (2004) SmartWeb Project.
26. Oberle, D.: Semantic Management of Middleware. Volume I of The Semantic Web and Beyond. Springer (2006)
27. Gangemi, A., Mika, P.: Understanding the semantic web through descriptions and situations. In: Databases and Applications of Semantics (ODBASE 2003), Catania, Italy (November 3–7 2003)
28. Sonntag, D.: Towards interaction ontologies for mobile devices accessing the semantic web - pattern languages for open domain information providing multimodal dialogue systems. In: Proceedings of the workshop on Artificial Intelligence in Mobile Systems (AIMS). 2005 at MobileHCI, Salzburg (2005)
29. Hovy, E., Gerber, L., Hermjakob, U., Lin, C.Y., Ravichandran, D.: Towards semantic-based answer pinpointing. In: Proceedings of Human Language Technologies Conference, San Diego CA. (March 2001) 339–345

30. Sonntag, D., Romanelli, M.: A multimodal result ontology for integrated semantic web dialogue applications. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy (May 24–26 2006)
31. Hunter, J.: Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In: Proceedings of the International Semantic Web Working Symposium (SWWS). (2001)
32. Benitez, A.B., Rising, H., Jorgensen, C., Leonardi, R., Bugatti, A., Hasida, K., Mehrotra, R., Tekalp, A.M., Ekin, A., Walker, T.: Semantics of Multimedia in MPEG-7. In: IEEE International Conference on Image Processing (ICIP). (2002)
33. Romanelli, M., Sonntag, D., Reithinger, N.: Connecting foundational ontologies with mpeg-7 ontologies for multimodal qa. In: Proceedings of the 1st International Conference on Semantics and digital Media Technology (SAMT), Athens, Greece (December 6-8 2006)
34. Ghallab, M., Nau, D., Traverso, P.: Automated planning. Elsevier Kaufmann, Amsterdam (2004)
35. Ankolekar, A., Hitzler, P., Lewen, H., Oberle, D., Studer, R.: Integrating semantic web services for mobile access. In: Proceedings of 3rd European Semantic Web Conference (ESWC 2006). (2006)
36. Engel, R.: Robust and efficient semantic parsing of free word order languages in spoken dialogue systems. In: Proceedings of 9th Conference on Speech Communication and technology, Lisboa (2005)
37. Gavaldà, M.: SOUP: A parser for real-world spontaneous speech. In: Proc. of 6th IWPT, Trento, Italy (February 2000)
38. Potamianos, A., Ammicht, E., Kuo, H.K.J.: Dialogue managment in the bell labs communicator system. In: Proc. of 6th ICSLP, Beijing, China (2000)
39. Ward, W.: Understanding spontaneous speech: the Phoenix system. In: Proc. of ICASSP-91. (1991)
40. Kaiser, E.C., Johnston, M., Heeman, P.A.: PROFER: Predictive, robust finite-state parsing for spoken language. In: Proc. of ICASSP-99. Volume 2., Phoenix, Arizona (1999) 629–632
41. Lavie, A.: GLR*: A robust parser for spontaneously spoken language. In: Proc. of ESSLLI-96 Workshop on Robust Parsing. (1996)
42. Huynh, D.T.: Communicative grammars: The complexity of uniform word problems. Information and Control **57**(1) (1983) 21–39
43. Becker, T.: Natural language generation with fully specified templates. In Wahlster, W., ed.: SmartKom: Foundations of Multi-modal Dialogue Systems. Springer, Heidelberg (2006) 401–410
44. Engel, R.: Spin: A semantic parser for spoken dialog systems. In: Proceedings of the 5th Slovenian First International Language Technology Conference (IS-LTC 2006). (2006)
45. Pfleger, N.: Fade - an integrated approach to multimodal fusion and discourse processing. In: Proceedings of the Dotoral Spotlight at ICMI 2005, Trento, Italy (2005)
46. Pfleger, N., Alexandersson, J.: Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems. In: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial), Postdam, Germany (September 11-13 2006) 2–9
47. Porzel, R., Zorn, H.P., Loos, B., Malaka, R.: Towards a Separation of Pragmatic Knowledge and Contextual Information. In: Proceedings of ECAI 06 Workshop on Contexts and Ontologies, Riva del Garda, Italy (2006)

48. Hacker, C., Batliner, A., Nöth, E.: Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In Sojka, P., Kopecek, I., Pala, K., eds.: Text, Speech and Dialogue. 9th International Conference (TSD 2006). Number 4188 in Lecture Notes in Artificial Intelligence (LNAI), Heidelberg, Germany, Springer (2006) 581–588
49. Matheson, C., Poesio, M., Traum, D.: Modelling grounding and discourse obligations using update rules. In: Proceedings of NAACL 2000. (May 2000)
50. Sonntag, D.: Towards combining finite-state, ontologies, and data driven approaches to dialogue management for multimodal question answering. In: Proceedings of the 5th Slovenian First International Language Technology Conference (IS-LTC 2006). (2006)
51. Carpenter, B.: The logic of typed feature structures (1992)
52. Larsson, S., Traum, D.: Information state and dialogue management in the TRINDI dialogue move engine toolkit. Natural Language Engineering, Cambridge University Press (2000)
53. Matheson, C., Poesio, M., Traum, D.: Modelling grounding and discourse obligations using update rules. In: Proceedings of NAACL 2000. (May 2000)
54. Walker, M., Fromer, J., Narayanan, S.: Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email (1998)
55. Singh, S., Litman, D., Kearns, M., Walker, M.: Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. Journal of Artificial Intelligence Research (JAIR), Volume 16, pages 105-133 (2002)
56. Rieser, V., Kruijff-Korbayova, K., Lemon, O.: A framework for learning multimodal clarification strategies. In: Proceedings of the International Conference on Multimodal Interfaces (ICMI). (2005)
57. Raghavan, H., Madani, O., Jones, R.: When will a human in the loop accelerate learning. In: Proceedings of the International Workshop on Artificial Intelligence for Human Computing at the International Joint Conference on Artificial Intelligence IJCAI'07, Hyderabad, India (2007) 97–105

A Learning-Based High-Level Human Computer Interface for Face Modeling and Animation

Volker Blanz

Institute for Vision and Graphics, Universität Siegen, Germany
blanz@informatik.uni-siegen.de
http://mi.informatik.uni-siegen.de

Abstract. This paper describes a system for animation and modeling of faces in images or in 3D. It provides high-level control of facial appearance to users, due to a learning-based approach that extracts class-specific information from a database of 3D scans. The modification tools include changes of facial attributes, such as body weight, masculine or feminine look, or overall head shape. Facial expressions are learned from examples and can be applied to new individuals. The system is intrinsically based on 3D face shapes and surface colors, but it can be applied to existing images as well, using a 3D shape reconstruction algorithm that operates on single images. After reconstruction, faces can be modified and drawn back into the original image, so the users can manipulate, animate and exchange faces in images at any given pose and illumination. The system can be used to create face models or images from a vague description or mental image, for example based on the recollection of eyewitnesses in forensic applications. For this specific problem, we present a software tool and a user study with a forensic artist. Our model-based approach may be considered a prototype implementation of a high-level user interface to control meaningful attributes in human faces.

1 Introduction

Unlike many other fields in computer science, progress in computer graphics involves adaptations to human users in a variety of different ways: First, the image data produced in graphics has to consider and exploit the laws of human perception. For example, research in tone mapping aims at producing images at low dynamic range in intensity, which can be displayed on standard computer screens, and still reproduce the visual appearance of natural scenes that have a large dynamic range due to extreme lighting conditions. Many algorithms in tone mapping are inspired by psychophysical findings (such as [14]), and in turn, the only valid criterion for the quality of a tone mapping operator is human perception [20].

Computer graphics is being adapted to the human needs also in terms of the contents and style of the images, by producing material that is meaningful in a

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 296–315, 2007.

cultural context. While most of the computer graphics images and movies produced in the 1980s and 1990s were entirely virtual scenes, located in a physically and semantically void space, visual effects today are more and more embedded into a rich context: Computer graphics is mixed with shots of natural scenes, with either of the two dominating the picture, and existing, real characters, images and objects are reproduced and manipulated. Unlike the artificial characters of early computer animation, they may now be used as virtual stunt doubles in movies such as *The Matrix*. Scenes may be altered by adding or removing buildings or other objects in each frame. The benefit of the combination of 3D graphics with natural images is a photorealistic, highly complex visual appearance with meaningful content. However, it is still challenging to achieve the same visual standard for computer graphics elements when shown side by side with natural photo material.

In facial animation, we are beginning to see a level of quality that captures even subtle facial expressions or combinations of expressions. These give artists the tools to model the complex, sometimes conflicting emotions of their characters. An example of this can be seen in the character *Gollum* in the feature film *The Lord of the Rings*. A medium that conveys more and more emotional content in artistically sophisticated narratives, computer graphics is transforming from a machine-centered to a human-centered medium, with the content not limited but enhanced by technology.

In contrast to the strive for photorealistic images, non-photo-realistic rendering has developed algorithms that give artistic styles to images and movies, some of them simulating painterly styles such as oil on canvas or watercolor, pen and ink drawings, cartoon drawings and engravings ([10,12]). Non-photorealistic rendering, therefore, bridges the gap between electronic and traditional art, and brings a variety of interesting connotations to otherwise synthetic images. Non-photorealistic rendering may also be a powerful tool for visualization, and for making image content more comprehensible to human viewers [6].

Finally, the progress in user interfaces for content creation in computer graphics adapts this tool more and more to the demands of the artist, and it remains an interesting challenge to make the increasingly powerful algorithms in graphics easily available to users with little or no technical background. Moreover, the efficiency of the tools provided by computer graphics systems may help to increase the creative power of those who use it significantly. An important component is the level of abstraction of the interaction tools and of the internal representations of computer graphics systems: As an example, consider an artist who wishes to make a character in an image more skinny. In an image-based system, such as software for digital image processing, the artist would need to shift the facial silhouette of the cheeks with copy-and-paste tools, or even paint the new silhouette with a digital brush. In 3D computer graphics, the face would be represented by a polygon mesh, and the artist would select and shift groups of vertices to change the 3D shape of the cheeks. On the highest level of abstraction, however, the artist would like to have an interactive tool, such as a slider,

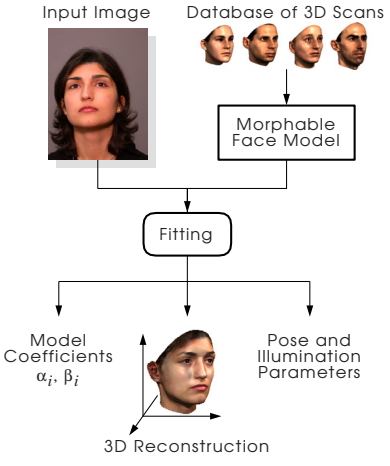


Fig. 1. Fitting the Morphable Model to an image produces not only a 3D reconstruction, but also model coefficients α_i, β_i, and an estimate of head orientation, position and illumination

that controls the skinniness or fatness of a face directly, so the user would only select a face in the image and use that slider.

In this paper, we describe a system that implements this paradigm of high-level control. The system works on faces in any given photo or painting at any pose and illumination. Applied on existing image material, it provides a tool that interacts with material that is meaningful in a complex cultural background, such as paintings. As a tool that animates, modifies or exchanges faces, it addresses perhaps the most relevant content of visual media, the human face.

1.1 System Overview

Modification of faces in given images at any pose and illumination is a non-trivial problem that involves information about the 3D geometry of the face. For example, making a face more fat or skinny changes the silhouette and the shading of the face, which calls for a direct or indirect (implicit) representation of effects in 3D space, such as perspective projection, occlusion and interaction of light with matter. In our approach, we chose an explicit representation of the 3D geometry of the face.

In order to interact with faces in 3D space, given an image, we have to solve the difficult, ill-posed problem of 3D shape reconstruction from single images. This problem can only be solved by including prior knowledge about the possible 3D solutions. In 3D reconstruction from architecture, such prior knowledge may be the fact that many lines are parallel or orthogonal in 3D [11]. For human faces, the set of 3D solutions may be restricted by exploiting the statistics of face shapes. The core of our work, therefore, is a *3D Morphable Model* of faces [4] that captures the natural variations observed in human faces. This model is

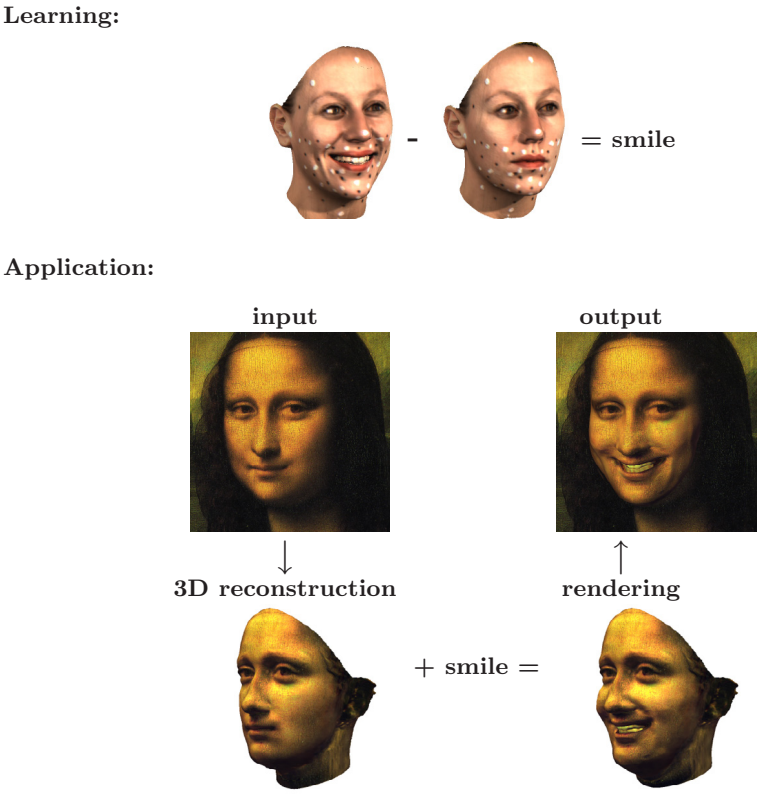


Fig. 2. In the vector space of faces, facial expressions are transferred by computing the difference between two scans of the same person (top row), and adding this to a neutral 3D face. To modify Leonardo's Mona Lisa (second row), we reconstruct her 3D face (third row), add the expression, and render the new surface into the painting (second row, right).

learned automatically from a dataset of 3D scans of faces [4]. Figure 1 illustrates the process of 3D shape reconstruction. The algorithm for shape reconstruction is fully automated, but it has to be initialized by manually labelling a set of between 6 and 15 feature points in the image. As a side effect, the algorithm computes the 3D pose and illumination parameters of the face in the scene, which can be used when the modified 3D face is drawn back into the input image.

For 3D face manipulation, we describe three modes of interaction:

- Facial animation,
- High-level control of facial attributes, such as gender, body weight and nose shape,
- Exchanging faces in images.

In all of these interactions, a 3D model of the face is reconstructed from the input image, modified in 3D, and drawn back into the original image (Figure 2.)

Both the animation and the modification of high-level attributes are learning-based: changes in 3D geometry and in texture are learned from datasets of 3D scans, unlike the labor-intensive manual surface editing procedures performed by animators and modelers in production studios today.

In the following sections, we describe the Morphable Model (Section 2), summarize the algorithm for 3D shape reconstruction (Section 3), describe the approach to facial animation (Section 4), discuss our technique for exchanging faces in images (Section 5), present a method for learning and changing attributes in faces (Section 6) and describe a prototype implementation of a user interface for forensic applications (Section 7).

2 A Morphable Model of 3D Faces

The Morphable Model of 3D faces [18,4] is a vector space of 3D shapes and textures spanned by a set of examples. Derived from textured *Cyberware* (TM) laser scans of 200 individuals, the Morphable Model captures the variations and the common properties found within this set. Shape and texture vectors are defined such that any linear combination of examples $\mathbf{S}_i$, $\mathbf{T}_i$,

$$\mathbf{S} = \sum_{i=1}^{m} a_i \mathbf{S}_i, \quad \mathbf{T} = \sum_{i=1}^{m} b_i \mathbf{T}_i. \tag{1}$$

is a realistic face if $\mathbf{S}$, $\mathbf{T}$ are within a few standard deviations from their averages. Each vector $\mathbf{S}_i$ is the 3D shape of a polygon mesh, stored in terms of x, y, z-coordinates of all vertices $j \in \{1, \ldots, n\}$, $n = 75972$:

$$\mathbf{S}_i = (x_1, y_1, z_1, x_2, \ldots, x_n, y_n, z_n)^T. \tag{2}$$

In the same way, we form texture vectors from the red, green, and blue values of all vertices' surface colors:

$$\mathbf{T}_i = (R_1, G_1, B_1, R_2, \ldots, R_n, G_n, B_n)^T. \tag{3}$$

Such a definition of shape and texture vectors is only meaningful if the vector components of all vectors $\mathbf{S}_i$, $\mathbf{T}_i$ have point-to-point correspondence: Let $x_{i,j}$,$y_{i,j}$,$z_{i,j}$, be the coordinates of vertex j of scan i. Then, for all scans i, this has to be the same point, such as the tip of the nose or the corner of the mouth. Using an algorithm derived from optical flow, we compute dense correspondence for all $n = 75972$ vertices automatically [4].

Finally, we perform a Principal Component Analysis (PCA, see [9]) to estimate the probability distributions of faces around their averages $\overline{\mathbf{s}}$ and $\overline{\mathbf{t}}$. This gives us a set of m orthogonal principal components $\mathbf{s}_i$, $\mathbf{t}_i$, and the standard deviations $\sigma_{S,i}$ and $\sigma_{T,i}$ of the dataset along these axes. We can now replace the basis vectors $\mathbf{S}_i$, $\mathbf{T}_i$ in Equation (1) by $\mathbf{s}_i$, $\mathbf{t}_i$:

$$\mathbf{S} = \overline{\mathbf{s}} + \sum_{i=1}^{m} \alpha_i \cdot \mathbf{s}_i, \qquad \mathbf{T} = \overline{\mathbf{t}} + \sum_{i=1}^{m} \beta_i \cdot \mathbf{t}_i. \tag{4}$$

In the following, we use the $m = 149$ most relevant principal components only, since the other components tend to contain noise and other non class-specific variations.

3 Estimation of 3D Shape, Texture, Pose and Lighting

From a given set of model parameters α and β (4), we can compute a color image $\mathbf{I}_{model}(x, y)$ by standard computer graphics procedures, including rigid transformation, perspective projection, computation of surface normals, Phong illumination, and rasterization. The image depends on a number of rendering parameters ρ. In our system, these are 22 variables:

- 3D rotation (3 angles)
- 3D translation (3 dimensions)
- focal length of the camera (1 variable)
- angle of directed light (2 angles)
- intensity of directed light (3 colors)
- intensity of ambient light (3 colors)
- color contrast (1 variable)
- gain in each color channel (3 variables)
- offset in each color channel (3 variables).

All parameters are estimated simultaneously in an analysis-by-synthesis loop. The main goal of the analysis is to find the parameters α, β, ρ that make the synthetic image $\mathbf{I}_{model}$ as similar as possible to the original image $\mathbf{I}_{input}$ in terms of pixel-wise image difference in the red, green and blue channel:

$$E_I = \sum_x \sum_y \sum_{c \in \{r,g,b\}} (I_{c,input}(x, y) - I_{c,model}(x, y))^2. \tag{5}$$

All scene parameters are recovered automatically, starting from a frontal pose in the center of the image, and at frontal illumination. To initialize the optimization process, we use a set of between 6 and 15 feature point coordinates [5]: The manually defined 2D feature points $(q_{x,j}, q_{y,j})$ and the image positions (p_{x,k_j}, p_{y,k_j}) of the corresponding vertices k_j define a function

$$E_F = \sum_j \| \begin{pmatrix} q_{x,j} \\ q_{x,j} \end{pmatrix} - \begin{pmatrix} p_{x,k_j} \\ p_{y,k_j} \end{pmatrix} \|^2. \tag{6}$$

that is minimized along with the image difference E_I in the first iterations.

In order to avoid overfitting effects that are well-known from regression and other statistical problems (see [9]), we add regularization terms to the cost function that penalize solutions that are far from the average in terms of shape, texture, or the rendering parameters:

$$E_{reg} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \overline{\rho}_i)^2}{\sigma_{R,i}^2}. \tag{7}$$

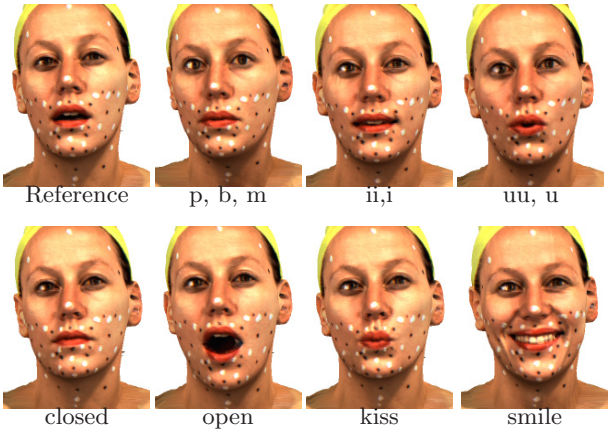


Fig. 3. Examples from the dataset of 35 static 3D laser scans forming the vector space of mouth shapes and facial expressions. 17 scans show different visemes, others show the mouth opening gradually.

$\overline{\boldsymbol{\rho}}_i$ are the standard starting values for ρ_i, and $\sigma_{R,i}$ are ad–hoc estimates of their standard deviations. The full cost function

$$E = \frac{1}{\sigma_I^2} E_I + \frac{1}{\sigma_F^2} E_F + E_{reg} \tag{8}$$

can be derived from a maximum-a-posteriori approach that maximizes the posterior probability of α, β and ρ, given $\mathbf{I}_{input}$ and the feature points [4,5]. The optimization is performed with a Stochastic Newton Algorithm [5]. The fitting process takes 3 minutes on a 3.4GHz Xeon processor.

The linear combination of textures $\mathbf{T}_i$ cannot reproduce all local characteristics of the novel face, such as moles or scars. We extract the person's true texture at high resolution from the image by an illumination-corrected texture extraction algorithm [4]. At the boundary between the extracted texture and the predicted regions, we produce a smooth transition based on a reliability criterion for texture extraction that depends on the angle between the viewing direction and the surface normal. Structures that are visible on one and occluded on the other side of the face can be reflected, due to facial symmetry.

4 Facial Animation in Images

The Morphable Model can not only represent variations across the faces of different persons, but also changes in shape and texture of an individual face due to facial expressions. Morphing between such face vectors generates smooth, continuous transitions as they occur over time t when a face moves:

$$\mathbf{S}(t) = (1 - \lambda(t)) \cdot \mathbf{S}_{expression1} + \lambda(t) \cdot \mathbf{S}_{expression2}, \tag{9}$$

Fig. 4. Reconstructed from the original images (left column), 3D shape can be modified automatically to form different mouth configurations. The paintings are Vermeer's "Girl with a Pearl Earring", Tischbein's Goethe, Raphael's St. Catherine, and Edward Hopper's self-portrait. The bottom left image is a digital photograph. The wrinkles are not caused by texture, but entirely due to illuminated surface deformations. In the bottom-right image, they are emphasized by more directed illumination. Teeth are transferred from 3D scans (Figure 3). The open mouth in Vermeer's painting was closed by our algorithm automatically by projecting the reconstructed shape vector on the subspace of neutral faces (top row, second image).

with a scalar function $\lambda(t)$ that controls the rate at which the face transforms from expression 1 to expression 2.

Unlike previous approaches to facial animation, such as parameterized models that are designed by artists (for an overview, see [13]) or models that simulate the physical properties of muscles and tissue [16], our technique relies entirely on observations of facial expressions on human faces. An automated algorithm learns how points on the surface move as a person acts or speaks, and these movements can be transferred to novel faces.

In order to learn the degrees of freedom of faces in facial expressions and speech from data, we recorded a set of 35 static laser scans of one person (Figure 3). 17 of the scans show different visemes, which are the basic mouth shapes that occur in human speech. Mouth movements and expressions learned from these scans can then be transferred to new individuals by a simple vector operation (Figure 2):

$$\Delta\mathbf{S}_{expression} = \mathbf{S}_{expression,person1} - \mathbf{S}_{neutral,person1} \tag{10}$$

$$\mathbf{S}_{expression,person2} = \mathbf{S}_{neutral,person2} + \Delta\mathbf{S}_{expression}. \tag{11}$$

This simple linear approach assumes that the deformations of faces are identical for all individuals, which is only an approximation of the individual expressions observed in real faces. A full investigation and a statistical analysis of the individual differences in expressions requires a large database of different persons' expressions [19]. One of the challenges in this approach is to extrapolate to novel, unknown faces, and to avoid overfitting and other statistical problems. Our direct transfer of expressions, therefore, is a safe guess at the price of missing some of the idiosyncrasies.

To convert the scan data into shape and texture vectors of the Morphable Model, it is essential to compute dense point-to-point correspondence between different expressions in the same way as between scans of individuals (Section 2). However, the large differences in geometry between open and closed mouth scans, and the fact that surface regions such as the teeth are visible in some mouth poses and occluded in others, make this problem significantly more difficult, and requires additional techniques as described in [2]. Unlike the face, which is morphed in a non-rigid way during animation, the teeth are rigid surface elements located behind the lips in 3D space. Their position is fixed with respect to the head (upper teeth) or the chin (lower teeth). Taken from a single, open mouth scan of the subject shown in Figure 3, these teeth can be inserted into novel 3D faces by a simple scale and translation operation, based on the positions of the corners of the mouth [2]. Figure 4 shows a set of examples. In paintings, the strokes of the brush are captured by the high-resolution texture extracted from the image.

5 Exchanging Faces in Images

The Morphable Model and the fitting procedure described in Section 3 achieve a full separation of parameters that are characteristic for an individual, i.e. shape

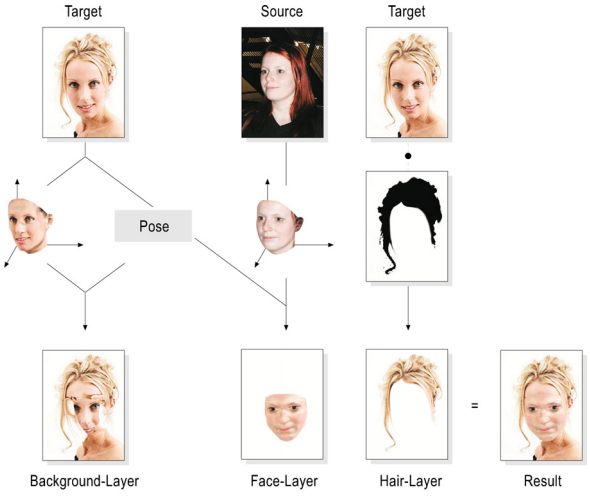


Fig. 5. For transferring a face from a source image (top, center) to a target (top, left), the algorithm builds a composite of three layers (bottom row): An estimate of 3D shape and scene parameters ("pose") from the target image helps to remove the contour of the original face in the background (left). The scene parameters of the target are also used to render the 3D face reconstructed from the source image (center column). A semi-automatic segmentation of hair defines the transparency mask for the hairstyle in the top layer (right column).

and texture, from scene parameters that are specific for a given image, such as pose and illumination. Therefore, it is straight forward to exchange faces in images by replacing the shape and texture parameters α β or vectors $\mathbf{S}$ and $\mathbf{T}$, while keeping the scene parameters unchanged [3]. This process is summarized in Figure 5.

The Morphable Model only covers the face, including forehead, ears and neck, but not the back of the head and the shoulder. We leave these unchanged by using the original image as a background to the novel face. Alpha blending along the cutting edges of the facial surface produces a smooth transition between original and modified regions. In contrast to cutting edges, the occluding contours along the silhouette of the face are not blended with the background, but drawn as sharp edges. When the novel face is smaller or more skinny than the original face, the original silhouette would be visible next to the contour of the 3D model. In order to prevent this from happening, we have proposed a simple algorithm that removes the original silhouette line from the background image. In that image, the algorithm reflects pixels from outside of the face across the silhouette to the inside (Figure 5, see [3] for details.) This can be done automatically, because the original contour line is known from the projection of the reconstructed 3D model of the original face.

In some images, strands of hair or other objects occlude part of the facial surface. For these cases, we have proposed a simple compositing algorithm. Based on pixel luminance and with some additional, manual interaction, an alpha map

for the foreground structures is created, and with this alpha map, the original image is composited on top of the new face (Figure 5.)

The algorithm has a number of applications in image processing, for example in consumer software for digital image processing, and in projects such as a virtual try-on for hairstyles (Figure 6.) If applied to image sequences, it could also be interesting for video post-processing.

6 Learning-Based Modification of High-Level Attributes

Interactive control of perceptually meaningful features of faces, such as skinny versus obese appearance, involves global changes of face shape and texture, unlike the local changes in pixel values or vertex positions in 2D image processing and 3D mesh editing, respectively. One way of providing such tools would be to collect a set of deformation patterns, similar to the facial expressions in Section 4, that are manually defined, and apply those to novel faces automatically within the framework of the Morphable Model. However, modeling such deformations would be a challenging task that requires careful observation of human anatomy, and artistic skill to implement these observations on a prototype face. Instead, we propose an approach that only involves human *ratings* of attributes, such as skinniness, rather than any kind of *description*. Rating is a much easier task for humans, and it can be performed even on the most subjective attributes, such as attractiveness of faces. Based on ratings of a set of example data (shape vectors or texture vectors of faces), we compute a vector that can be added to or subtracted from a given face to change the attribute, while all other attributes and the individual characteristics of the face are left unchanged.

Let $\boldsymbol{x}_i$, $i = 1...m$, be a set of sample vectors (shape or texture), and $b_i \in I\!R$ be the ratings of a given attribute for these 3D faces. b_i can be either given as ground truth, e.g. for gender, or based on subjective ratings by the user. Our approach is to estimate a function f that assigns attribute values to faces, and then follow the gradient of this function in order to achieve a given change in attribute at a minimal, most plausible change in appearance [4,1]. Given the limited set of data, we choose a linear regression for f, and minimize the least squares error

$$E = \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - b_i)^2. \tag{12}$$

It can be shown that the gradient $\boldsymbol{a}$ of the optimal function f depends on an appropriate definition of the distance measure in face space. A perceptually meaningful distance measure is given by the probability distribution of faces in terms of PCA [1]. Then, the optimal vector for changing attributes turns out to be a simple weighted sum of the example data (for details, see [1]):

$$\boldsymbol{a} = \frac{1}{m} \sum_{i=1}^{m} b_i \boldsymbol{x}_i. \tag{13}$$

Adding multiples $\boldsymbol{x} \mapsto \boldsymbol{x} + \lambda \boldsymbol{a}$, with $\lambda \in I\!R$, will change facial attributes in the desired way and leave all characteristics that are uncorrelated with the attribute

Fig. 6. In a virtual try-on for hairstyles, the customer (top, left) provides a photograph, and selects target images of hairstyles (top, right). Then, the system creates synthetic images of the customers face pasted into these hairstyles.

unchanged. We would like to point out that the linear attribute function f is only a first order approximation of the correct, non-linear function. Still, our results indicate that for many attributes, this approximation produces realistic results on 3D scans or on faces that were reconstructed from images (Figure 7)

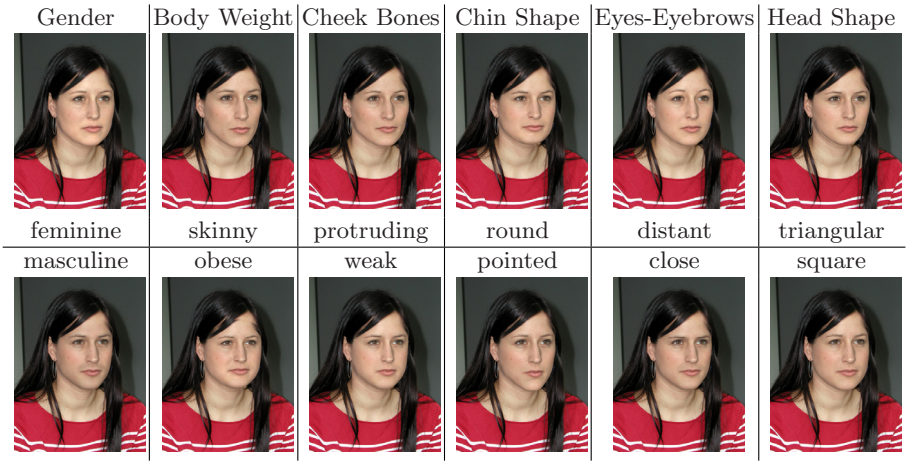


Fig. 7. Learned from labeled examples, facial attributes can be manipulated in 3D faces automatically, and in combination with 3D shape reconstructions, manipulations can be performed within given portraits at any pose and illumination. Attributes in regions such as the eyes are treated as global shape vectors, which reveals correlations between different regions of the face observed in the data. For example, the distance between eyes and eyebrows turns out to be correlated with chin shape (fifth column). This correlation can be suppressed by setting vector components of $\boldsymbol{a}$ to zero outside of the region of interest, which makes the changes local.

7 A Human Computer Interface for Face Modeling

The technology that was described in tbe previous sections can be integrated into an advanced user interface that operates on a high level of abstraction in two respects: (1) Faces are manipulated in a way that is semantically meaningful to the user, such as facial expression or facial attributes, (2) all operations are applied to a 3D model, but can be projected into existing images at any time, which makes the system very flexible, but still retains much of the photorealism of image-based techniques.

We focus on the problem of creating pictures of faces from mental images or vague ideas about the desired appearance of faces. This has important applications in both art/design and forensic sciences: new virtual characters for movies, advertisement or computer games are created from the artist's imagination, and illustrations of criminals need to be generated from the eyewitnesses' descriptions. In both cases, the artist/witness (denoted as *source* in the following) has a mental image of the *target face* to be created, and it is important not to influence this mental image during the process of reconstructing the face. In general, the sources cannot give a complete verbal description of the target face, and they tend to recollect only some striking details, for instance bushy eyebrows or a hooked nose, while they are unaware of many other characteristics. In the future, we hope that the system described in this section could substitute

commercial systems used by police for creating composite or photofit pictures of suspects at higher quality and with considerably less effort.

In contrast to most of these commercial systems, our approach has the following advantages:

- Our GUI offers intuitive ways to modify features of the target face in arbitrary order.
- Unspecified parts of the reconstructed face are automatically completed according to statistical properties.
- Anatomical/ethnical correlations within a face are taken into account automatically during reconstruction.
- A 3D face model is created that can be viewed from an arbitrary view point with arbitrary lighting conditions.
- The resulting face model can be rendered automatically into background images in appropriate pose and illumination.

7.1 Components of the Modeling Interface

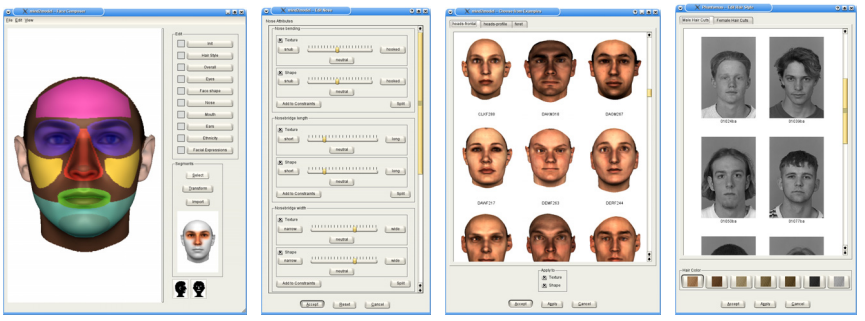
The system operates on 3D faces, but the user can project the face into an existing image at any time to see how it looks in a given setting. Starting from the average 3D face of the Morphable Model, the user can change the appearance interactively in a range of different ways according to the Morphable Model technique, while some other modes of interaction are still inspired by the traditional photofit paradigm. In a user study, however, we found that the new, attribute-based interaction paradigm was more appropriate to implement the input that was given by witnesses [1].

7.2 General Settings

Before even being shown any of the faces, the user (i.e., either the source or an operator) may specify age, gender, and ethnicity of the target face. Setting these parameters modifies the starting head and restricts the selection of example faces shown in the import dialog box to match the criteria specified for age, gender, and ethnicity (see Section 7.6). The general settings help to reduce the exposure of witnesses to face images, and thus avoid the confusing interaction between the image material with the mental images. Witnesses often report being distracted by the fact that they see a large number of faces in traditional systems, an observation that is in part supported by psychological studies.

7.3 Simultaneous Versus Separate Modification of Image Regions

Various editing operations on individual facial regions (such as the nose, eyes, eyebrows, mouth) may affect the whole face due to correlations among the features and the overall face shape that we have derived from the 3D dataset. For instance, bushy eyebrows are usually correlated to a male face and to a darker skin, whereas a wide mouth is typically correlated to a broader face and also to a broader nose. It is one of the main advantages of our approach that these



(a) (b) (c) (d)

Fig. 8. Snapshots of our system's GUI. Left to right: (a) Main window in selection mode. The user clicks on one or several segments (color-coded in the main drawing area), which are then added to the selection mask shown in the small pixmap on the right side (see also Section 7.3). In this example, the user has selected both eyes and the nose for further editing operations. (b) Dialog for setting individual facial attributes of the nose. See Section 7.4 for details. (c) Database dialog for selecting features from a large database of example faces (see Section 7.6). (d) In the hair style dialog, the user selects a hair style and color (see Section 7.7).

correlations are taken into account automatically. In cases where only little is known or remembered of the target face, exploiting these correlations may add significantly to the faithfulness of the reconstruction. On the other hand, if the user knows exactly what detail should be changed without influencing any other facial attributes, it is desirable to constrain the effect of editing operations to a local area.

In our system, the user can select one or several face *segments*, which define the area on the face where editing operations may have an effect. Figure 8(a) shows the GUI of our program in the selection mode, where the different segments are color-coded. The user clicks on one or several different segments, which are then added to the selection mask shown in the small pixmap on the right. Some facial features are represented by a hierarchy of segments, which are color-coded in different shades of the same base-color. The nose, for instance, consists of individual segments for the nose bridge, the alar wings, the tip, and the overall nose area. In the selection mode, the user may add segments from any hierarchy level of any facial feature to the selection mask. To ensure a smooth transition at the border of segments, we apply a multiresolution decomposition and blending technique [7].

7.4 Facial Attributes

The main interaction paradigm in our system is based on the high-level attributes described in Section 6. They affect both shape and/or color of facial features and can be applied to shape and color separately or to both at the same time. The facial attributes are organized in groups according to the facial feature they

affect. For example, all facial attributes that affect the shape or color of the nose are grouped together. Facial attributes can be modified on a continuous scale from $[-1 \ldots 1]$ (where 0 denotes the default value of the average face) using sliders. Figure 8(b) shows a dialog of our GUI with sliders for a variety of facial attributes of the nose. As a first step, we have included the following facial attributes in our system:

overall	: masculine–feminine, slender–obese, dark–light skin, younger–older, intensity of freckles, intensity of beard shadow, attractiveness, caricature level
face shape	: round–angular, narrow–broad, pear–heart shape, hollow–puffy cheeks, prominence of cheek bones, pointed–broad chin, retruded–protruded chin, distance between lips and chin, intensity of double chin, intensity of nasolabial fold
eyes	: slitted–round, upwards–downwards inclination, horizontal distance of eyeballs, dark–light iris, color of iris, dark–light eyebrows, thin–bushy eyebrows, straight–curved eyebrows (separate for left and right side), horizontal distance of eyebrows, distance between eyebrows and eyes
nose	: short–long nose bridge, narrow–wide nose bridge, narrow–wide alar wings, flat–round alar wings, snub–hooked nose, distance between nose and mouth
mouth	: narrow–wide, thin–full lips, dark–light lip color, convex–concave lip line
ears	: small–large, flat–jug ears
ethnicity	: Caucasian, Asian, African
expressions	: smiling, angry, surprised, scared, deranged, disgusted

In addition, our system features a *caricature level* control that is useful to overstate all characteristic attributes of a reconstructed face. Psychological findings [8] have shown that making faces more typical by caricaturing techniques increases subjects' ability to recognize faces. Caricaturing is achieved by morphing a 3D face away from the average [4].

As explained in Section 7.3, some facial attributes are correlated to other facial features. Although this is a desired feature of our system, it might happen that the user wants a specific value for a facial attribute A to remain untouched when performing further editing operations on other facial attributes. This can be achieved by adding the current setting of attribute A to a list of constraints. Any subsequent modifications of the face will then be performed in a way that maintains the pre-set level of the attribute: For example, if the user has chosen to constrain the face to a given masculinity level and then lifts the eyebrows relative to the eyes (which makes a face more female, see column 5 in Figure 7), then the system will automatically enhance other male traits by adding more of the attribute vector of masculine faces. This is achieved by solving a linear system of equations that restricts the output face to the subspace of vectors with constant, user-defined attribute levels [1].

7.5 Affine Transformations

At any time during the reconstruction of the target face, affine transformations (rotation, translation, non-uniform scaling) may be applied to the entire face or to a currently selected segment. These transformations are useful, for instance, to adjust the position and orientation of the face to fit into a background image or to account for an asymmetric layout of facial features.

7.6 Importing Features from a Database

The user may select the desired feature(s) from a database of example faces as shown in Figure 8(c). If a selection mask has been defined previously, only the corresponding features are imported from the selected face in the database. For example, the user may find the face with the most appropriate nose among the example faces, and that nose will be inserted seamlessly into the target face and can be further edited if needed. As with attribute editing, the import of features from a database can be restricted to the shape and/or texture of the feature.

7.7 Hair Style

The 3D Morphable Model [4] is based on 3D scans that did not represent hair styles. Therefore, we cannot handle hair in the same way as other facial attributes. Despite impressive progress in recent years, 3D hair modeling and rendering is still an open field of research. Even with photorealistic rendering techniques, collecting a database of hair styles would require to re-model existing styles. Therefore, we follow the approach of Section 5 and render the 3D face model into photographs of hair styles taken from different viewpoints.

For our hair style database, we selected a frontal and a profile image of people with several different hair styles from the FERET database [15]. After manually selecting 5–15 feature points in each hair style image once, the algorithm automatically estimates pose, illumination, and contrast for inserting arbitrary target faces. This estimation works for both color and monochrome photographs. We have chosen the monochrome images from the FERET database, since it

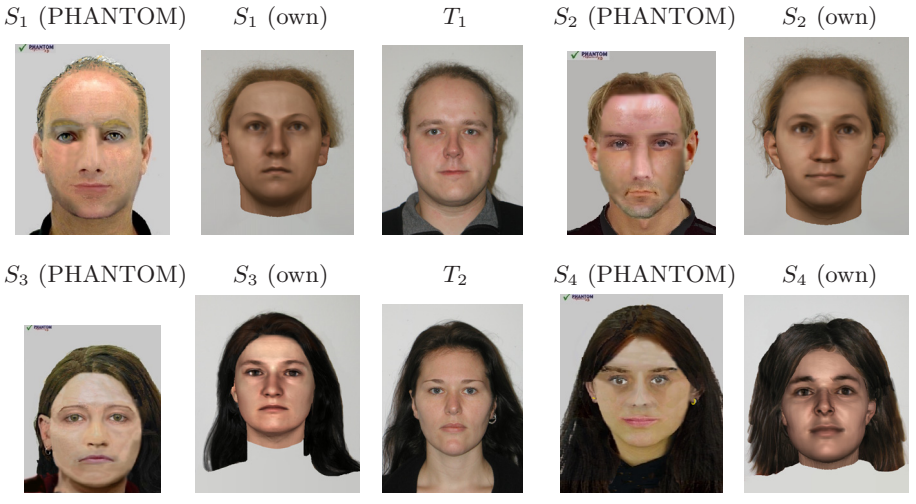


Fig. 9. Results of user study. Two target persons (T_1 and T_2) have been shown face-to-face to four source persons (S_1–S_4) for a duration of 60 seconds. The sources described the target faces after a delay of half-a-day up to one day. A professional forensic artist used the system [17] and Adobe Photoshop® for post-processing to create the composite images indicated by the keyword PHANTOM in about 2–3 hours each. The images created with our own approach took about 1.5–2 hours each.

was easy to color the hair in these images in seven different colors (different shades of blond, brown, red; black; grey) using an image manipulation program. Figure 8(d) depicts the hair style dialog of our system.

7.8 Exporting the Target Face

The target face can be saved as a 3D model for further operations. In particular, it is possible to change individual facial attributes later on or to display the face with different rendering parameters. In addition, the target face can be exported into an image using any desired set of rendering parameters. Rendering a face into an image containing scenery may help the source to recall details.

7.9 Results

In a user study, we have simulated the scenario of an eyewitness who has to create a facial composite image after a short encounter with a suspect [1]. Four sources (witnesses) saw two target persons (suspects) for 60 seconds in a face-to-face meeting. After a delay of half a day or one day, two of the sources were asked to help us create composite faces with our facial modeling software, supported by two operators who had not seen the target faces before. For comparison, a forensic artist made composite images with a commercial software system with the other two sources. Results (Figure 9) indicate that the 3D composite faces

generated by our system are comparable in quality to the results of the commercial system, but that they are obtained in a less tedious and time consuming way: With our system, it took 1.5 to 2 hours, while the traditional procedure took 2 to 3 hours.

Given the emotionally stressful situation of crime witnesses, we consider it an important contribution to reduce the pain in the production process of composite faces. Our high-level interface and the statistical analysis of semantically meaningful facial attributes may help to achieve this goal.

8 Conclusion

We have presented a framework for high-level manipulation of faces which can be applied both to 3D models and to existing image material. Our approach exploits prior knowledge about classes of objects that is learned automatically from a dataset of examples. We have shown results only for the class of human faces, but it can be extended to other relevant classes of objects, such as full human bodies and animals or perhaps even vehicles or buildings. When applied to images, the approach is very versatile in terms of the imaging conditions due to the internal 3D representation of the object class. In a case study, we have shown that our integrated system can help crime witnesses to create composite faces of suspects in a less tedious and painful way. This is due to the fact that the user interface of our system is adapted to the perception and language of humans, by learning and reproducing the meaningful attributes from a database of labeled samples. In Computer Graphics, we believe that the link between image-based and 3D-based image processing is a promising strategy to pursue in the future. The overall system, as it has been described in this paper, is an example of how computer graphics is adapted to humans in terms of the content that is displayed, but also with respect to the requirements of the users and the preferences of viewers.

Acknowledgments

The author would like to thank Thomas Vetter, Curzio Basso, Barbara Knappmeyer, Niko Troje, Heinrich Bülthoff, Tomaso Poggio, Kristina Scherbaum, Irene Albrecht, Jörg Haber and Hans-Peter Seidel for their collaboration in the projects summarized in this paper.

References

1. V. Blanz, I. Albrecht, J. Haber, and H.-P. Seidel. Creating face models from vague mental images. In E. Gröller and L. Szirmay-Kalos, editors, *Computer Graphics Forum, Vol. 25, No. 3 EUROGRAPHICS 2006*, pages 645–654, Vienna, Austria, 2006.
2. V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In P. Brunet and D. Fellner, editors, *Computer Graphics Forum, Vol. 22, No. 3 EUROGRAPHICS 2003*, pages 641–650, Granada, Spain, 2003.

3. V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In M.-P. Cani and M. Slater, editors, *Computer Graphics Forum, Vol. 23, No. 3 EUROGRAPHICS 2004*, pages 669–676, Grenoble, France, 2004.
4. V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Computer Graphics Proc. SIGGRAPH'99*, pages 187–194, 1999.
5. V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063– 1074, 2003.
6. M. Burns, J. Klawe, S. Rusinkiewicz, A. Finkelstein, and D. DeCarlo. Line drawings from volume data. In *Computer Graphics Proceedings SIGGRAPH 2005*, pages 512–518, 2005.
7. P. J. Burt and E. H. Adelson. A Multiresolution Spline with Application to Image Mosaics. *ACM Transactions on Graphics*, 2(4):217–236, Oct. 1983.
8. K. Deffenbacher, J. Johanson, T. Vetter, and A. O'Toole. The face typicality-recognizability relationship: encoding or retrieval locus? *Memory and Cognition*, 28(7):1173–1182, 2000.
9. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2nd edition, 2001.
10. A. Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Computer Graphics Proceedings SIGGRAPH 1998*, pages 453–460, 1998.
11. D. Liebowitz, A. Criminisi, and A. Zisserman. Creating architectural models from images. In *Proc. of EuroGraphics*, volume 18, pages 39–50, 1999.
12. V. Ostromoukhov. Digital facial engraving. In *Computer Graphics Proceedings SIGGRAPH 1999*, pages 417–424, 1999.
13. F. I. Parke and K. Waters. *Computer Facial Animation*. A K Peters, Wellesley, Massachusetts, 1996.
14. S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg. Time-dependent visual adaptation for fast realistic image display. In *Computer Graphics Proceedings SIGGRAPH 2000*, pages 47–54, 2000.
15. P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J*, 16(5):295–306, 1998.
16. D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.
17. UNIDAS. PHANTOM PROFESSIONALxp©. `http://www.unidas.com/html/phantome.html`, 2005.
18. T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.
19. D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *Computer Graphics Proc. SIGGRAPH'05*, pages 426–433, 2005.
20. A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel. Perceptual evaluation of tone mapping operators with real-world sceness. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)*, volume 5666 of *SPIE Proceedings Series*, pages 192–203, San Jose, USA, January 2005.

Challenges for Virtual Humans in Human Computing

Dennis Reidsma, Zsófia Ruttkay, and Anton Nijholt

Human Media Interaction, University of Twente,
PO Box 217, 7500AE Enschede, The Netherlands
{d.reidsma,z.m.ruttkay,a.nijholt}@ewi.utwente.nl
http://hmi.ewi.utwente.nl/

1 Introduction

The vision of Ambient Intelligence (AmI) presumes a plethora of embedded services and devices that all endeavor to support humans in their daily activities as unobtrusively as possible. Hardware gets distributed throughout the environment, occupying even the fabric of our clothing. The environment is equipped with a diversity of sensors, the information of which can be accessed from all over the AmI network. Individual services are distributed over hardware, share sensors with other services and are generally detached from the traditional single-access-point computer (see also the paper of Pantic *et al.* in this volume [51]).

'Unobtrusive support' means that where possible the user should be freed from the necessity of entering into an explicit dialog with all these services and devices. The environment shifts towards the use of *implicit interaction*, that is, "interactions that may occur without the behest or awareness of the user" [36]. However, not *all* interactions between user and environment will be implicit. It may not be possible, or it may not be desirable, e.g. because the user does not want to feel a loss of control over certain aspects of his environment. So how does the user achieve the necessary explicit interaction? Will (s)he address every query for information to the specific device or service that ultimately provides the information? Will (s)he give commands to the heating system, the blinds and the room lighting separately? Will each service and each device carry its own interaction interface? Clearly not. Interfaces will come to be developed that abstract from individual services and devices and offer the user access to certain combined functionalities of the system. The interfaces should support the mixture of explicit and implicit, and reactive and proactive, interaction required for a successful AmI environment. Finally, AmI environments are inherently multi-user, so the interface needs to be able to understand and engage in addressed multiparty interaction [48]. We argue that Virtual Humans (VHs) are eminently suited to fullfill the role of such interfaces.

An AmI environment can serve various purposes. It can be a home environment, an office environment, a public space or it can be used in an educational setting. Virtual humans can be available, among others, as friend, exercise adviser, health care specialist, butler, conversation partner or tutor. Sometimes they know things better than you do, sometimes they have more control over

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 316–338, 2007.

parts of the AmI environment than you have and sometimes they persuade you to do things differently. You may not always like all aspects of the virtual humans that cohabit your house. Maybe the virtual tutor that is available to monitor your children's homework sometimes takes decisions that are not liked by your children at all. Your virtual friend is not very interesting if it always agrees with your opininions. A healthcare agent has to be strict. A virtual human that acts as a conversational partner for your grandmother may have some peculiar behavior sometimes (like a dog or cat has; remember the Tamagotchi). As in collaborative virtual environments we can have remote participation in activities in AmI environments. Virtual humans can then represent family members (with all their characteristics, including weaknesses) that are abroad and that nevertheless take part in family activities. Transformation of communicative behavior of virtual humans that represent real humans can be useful too [4]. Summarizing, in the AmI environments we foresee that virtual humans can play human-like roles and need human-like properties, including (semi-) autonomous behavior, personalities, individual characteristics and peculiarities.

However, the vast majority of existing, implemented applications of Virtual Humans are focused around one clear task, such as selling tickets, learning a skill, answering questions or booking flights and hotels, and are accessed in a clearly explicit manner. There, one can take it as a given that the attention of the user is on the system and the user is primarily engaged with the interaction. In an AmI environment this is no longer true. A dialog with a Virtual Human may be altogether secondary to several other activities of the user. A dialog with a Virtual Human may also be about many different issues, pertaining to different aspects of the environment, in parallel. This has a lot of impact on many aspects of a Virtual Human.

In the rest of this paper we will examine a (not neccesarily exhaustive) number of aspects to Virtual Humans that we feel as most relevant to their introduction in a complex AmI environment. Some of these aspects relate to the embedding of the Human / Virtual Human interaction in ongoing daily activities: issues of synchronization, turn taking and control. Other points touch upon the fictional/real polemic: how realistic should VHs be? Should a VH interface exhibit 'socially inspired' behaviour? Should a VH exhibit also the imperfections and shortcomings so characteristic of human communication? Some of the points will be illustrated with examples from our recent work on Virtual Humans, summarized in Section 4.

1.1 Structure of This Paper

Different parts of this paper have appeared earlier in contributions to the workshops on Human Computing at the ICMI [64] and the IJCAI [65]. This paper supercedes and significantly extends the earlier publications. The paper is structured as follows. Section 2 sketches the background of Human Computing as we see it. Section 3 discusses the state of the art on Virtual Human research. Section 4 presents three example applications that have been developed at the Human Media Interaction (HMI) group, that will be used to illustrate points

throughout the paper. Subsequently, a number of aspects of VHs are discussed. Most of the aspects are centered around the general question of how human VHs really should be, which is raised in Section 5 and further elaborated in Sections 6 through 9. The paper ends with a short discussion in Section 10.

2 Human Computing: Background

Applications of computing technology are largely determined by their capabilities for *perception* of user and environment and for *presentation* of information to the user. An important characteristic of Human Computing is that the perception goes beyond the observation of superficial actions and events towards progressively more complex layers of *interpretation* of what is observed. Starting from observable events (form), such as gestures or speech, the system uses models of human interaction to interpret the events in terms of content, such as intentions, attitude, affect, goals and desires of the user. Elsewhere in this volume, the means of interpreting the users' behaviour are classified as the front end of Human Computing [51]. Conversely, in the other side of the system, which might then be called the back end, the intentions and goals of the system are realised in concrete actions, which have some direct effect on the environment and/or involve communication to the user.

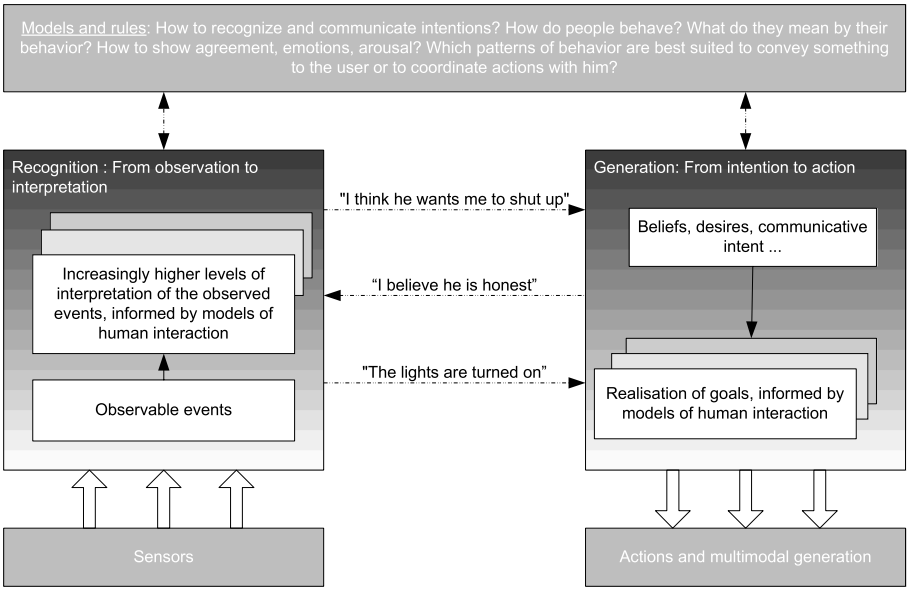


Fig. 1. Schematic overview of the various elements in the front-end and back-end of human computing, from observations via interpretations to appropriate reactions, mediated by various models expressing knowledge about the behavior of people

The models of how humans behave and communicate, mentioned above, are an integral part of the Human Computing paradigm. Not only for an adequate interpretation of the perceptions, but also to inform and inspire the patterns of interaction that should occur between system and user (see also [36]), and, especially when Virtual Humans inhabit the environment, as a basis for the realisation of the communicative intentions of the system. The concept is illustrated in Figure 1 and discussed in more detail in [58] and [60].

3 Virtual Humans: Related Work and State of the Art

Virtual Humans (VHs) [53] - also known under other names like humanoids [70], embodied conversational agents [11] or intelligent virtual agents - are computer models resembling humans in their bodily look and their communication capabilities. A VH can speak, display facial expressions, gaze and nod, use hand gestures and also perform motions of biological necessity, such as blinking or changing posture. Besides these 'back-end' related channels, a VH ideally should be equally capable in understanding speech and perceiving non-verbal signals of his human conversant partner. In generating and interpreting bodily signals, the VH should be mindful, to orchestrate his modality usage according goals, the state of the environment and last but not least, social communication protocols. From the point of view of HCI, there are two major motivation of 'putting a VH on the screen':

- By replacing the traditional computer-specific interaction modalities (keyboard, mouse) by the natural communicational capabilities of people, the services of computers will be accessible to a broad population, without respect to (computer) literacy, cultural and social background;
- VHs make new applications possible, where they fulfill traditional human roles, such as tutor, salesperson, and partner to play a game or chat with. Moreover, there are applications where the VH is in an entirely new role, without parallel in real life, such as human-like characters with fictional capabilities in (educational) games [28], interactive drama [43] or Virtual Reality applications [1], or a VH taking partial or split role known from real-life situations [6].

Often it is not easy to separate the two aspects in an application. For instance, a VH explaining how to operate a device [3], can be seen as a friendly interface replacing queries and search in an on-line textual help, but he also extends the services, by being able to demonstrate operations, possibly in a way tailored to the user's capabilities (e.g. handedness).

A key issue concerning VHs is the question whether they do indeed 'work'. In today's state of the art applications they cannot be mistaken for a real person. It is clear that they exist in a digital form, residing on a PC screen or projected to a wall as a set of pixels. Moreover, due to features in appearance and communication, they are not considered as a recording of a real person. The 'betraying features' may result from a conscious design decision to create a fictional character with limited human intelligence [38], or - more often - from technological

limitations such as not having fast enough models to simulate hair, or Text-To-Speech engines capable of producing synthetic speech with intonation and other meta-speech characteristics. In spite of these observations, there are two basic principles which make VHs promising as effective and engaging communicating partners of real people.

1. In a broader context, it was shown that humans treat computers as social actors [56]. We tend to talk about (and to) computers as living creatures, ask them to perform some task, get emotional about their performance and provide feedback accordingly. One may expect that it is even more so if we see a human-like creature representing the services of the underlying system.
2. With the appearance of a human-like character on the screen, all the subtleties (or lack of them) reminiscent of human communication will be interpreted and reacted to. Though we know that we are not facing a real person, we "suspend our disbelieve" against the VH: we do not look at them as cool computer engineering achievements, but react to them as we do to real people. This does not require, per se, photorealism in appearance and behavior. The very term was coined in connection with successful Disney animation figures, who thank their success to consistent behavior, both concerning changes in their emotions and goals reflecting some personality and in the way how this is conveyed in their speech and body language. Expressivity is achieved by using subtle phenomena of human communication in an enhanced or exaggerated - thus non-realistic - way.

Human-like appearance and behavior of the interface, in particular when the computer is hidden behind a virtual human, indeed strengthen the acceptance of the computer as a social actor [23] or even a persuasive social actor [22] that changes our attitudes and behavior.

But what design rules are to be followed making a 'successful' VH? One whom the users will like, find easy to communicate with... and who helps the user to perform the task required? The quality of the signals of communications modalities used by the VH should be good: his speech intelligible, his motion not jerky, his face capable of expressing emotions. In the past decade, much effort has been spent on improving the human-likeness of individual modalities of VHs, such as improving the quality of synthesized speech [45], modeling expressive gesturing of humans [30; 31], deriving computational models to capture the kinematics [72], providing means to fine-tune the effort and shape characteristics of facial expressions and hand gestures [14], model gaze and head behavior and adding biological motions like blinking or idle body motion [20]. The fusion of multiple modalities has been dealt with, from the point of view of timing of the generated behavior, and the added value of using multiple modalities in a redundant way.

Besides these low-level technical requirements, several subtle human qualities have been proven to play an essential role in both the subjective judgment of a (virtual) human partner and the objective influence of him. It has turned out that by using smalltalk, common ground can be created and as a consequence, the user will trust the VH more [8]. In addition to task-related feedback, showing interest and providing empathic reactions [19] contribute to the success of task

performance e.g. in tutoring [46; 41] and coaching. Emotional reactions are reminiscent of humans. Modeling emotions and their benefits in judging the VHs have been extensively addressed in recent works [25; 42]. Initially, the 6 basic emotions were to be shown on the face [21], which has been followed by research on the display of other emotions and of cognitive states, taking into account display rules regulating when emotions are to be hidden [54] or overcast by fake expressions e.g. to hide lies [57], studying principles to show mixed emotions on the face and to reflect emotions in gesturing [50]. Besides temporary emotions, having long-term attitude towards a VH, like friendship, has been pointed out [69].

Moreover,people attribute some personality to VHs based on signals of nonverbal communication, the wording of speech and physical impressions of the embodiment. It has been shown that signals of modalities such as posture or the intonation of synthetic speech are sufficient to endow VHs with a personality [47]. From these experiments it was also clear that people tend to prefer a virtual communicative partner with similar (or in some cases, complementary) personality. So an effective VH should be endowed with a personality best matching the envisioned user. Thus no single universal, best VH may be designed, but characteristics of the user - such as personality, but also age, ethnicity, gender - need to be taken into account. The importance of cultural and social connotation of a VH has been pointed out [52; 55].

It has also been mentioned that virtual humans should be individuals, in their behavior, verbal and non-verbal style, body, face and outfit [62]. Some subtle aspects of interaction, like the style of the VH [63] and his/her attitude towards the user have proven to be important in judging them.

As of the less investigated bodily design, some case studies show that similar criteria of judgment are to be taken into account when deciding about gender, age, ethnicity, formal or casual dressing [44], or even (dis)similarity to the human user [5] or non-stereotypical embodiment like a fat health-advisor [71]. Other dimensions of VH design are to do with human-likeness and realism. Should the VH be humanlike, or are animals or even objects endowed with human-like talking/gesturing behavior more appropriate? The application context may be decisive: for children, it may be a factor of extra engagement when the tutor is a beloved animal known from a favorite tale (or TV program), or when an animal or object to be learned about is capable of telling informative facts about 'himself'. But non-human characters may be beneficial for adult users too, to suggest the limited mental capabilities of the VH (using e.g. a dog [35], or a paperclip as Microsoft did). The degree of realism of the VH is another dimension. Most of the current research has been driven by the motivation to be able to reproduce human look and behavior faithfully. Recently, the expressivity in non-photorealistic and artistic rendering [61] and in non-realistic communication as done in traditional animation films [74; 12] and on theater stage [68] are getting a place in VH design.

The overall objectives of likability, ease of use and effectivity may pose conflicting design requirements. In an early study it was shown that even merely a slightly frowned eyebrow from a virtual human can result in measurable

differences: a stern face was more effective, but liked less, than a neutral face [73]. There are some applications like entertaining games or crisis management where basically only one of the two objectives is to be met. However, in most of the applications both factors play a role, and the two objectives need to be balanced with respect to the progress made and the current emotional and cognitive state of the user.

The potentials of VHs are huge. There are still major challenges to specific disciplines to improve bodily and mental capabilities of VHs along the above dimensions, and to compile these components into full-fledged, consistent believable virtual characters [26]. Also, the necessity of cooperation of different disciplines, and particularly, dedicated studies providing basis for computational models of human-human interaction, have been underlined [34]. In such a 'big scenario' with yet many problems to be solved, many more subtle issues of human - (virtual) human communication are not (yet) looked at. In the rest of the paper, we focus on such issues. They are essential to have VHs in an AmI environment, where they are active and present in the daily life of the users. On the other hand, several of these subtle issues have major consequences on the design of VHs and raise further, principal questions concerning the VHs. This is the major focus of the forthcoming sections. In this paper we do not address the technical feasibility of the envisioned characteristics, as each of them would require in-depth expertise from specific fields. However, we note that several of the features we discuss would be feasible with present-day technology, while others would need research in new directions in some disciplines both to provide computational models and good enough implementations.

4 Virtual Humans: Engagement and Enjoyment

4.1 Introduction

In this section, we present three applications currently being developed at the HMI (Human Media Interaction) research group: the Virtual Dancer [59], the Virtual Conductor [9] and the Virtual Trainer [67]. These three novel applications are summarized in preparation to our general discussion on subtleties of VHs, where we will use illustrative examples from the applications. All three applications require virtual humans with capabilities beyond the ones in more restricted or traditional functions such as providing information or tutoring. These seemingly very different applications share some basic features, and have actually been developed relying on a similar framework. In all three applications, the VH:

- has visual and acoustic perception capabilities,
- has to monitor and react to the user continuously,
- has to use subtle variants of a motion repertoire generated on the fly, and
- uses both acoustic (music, speech) and nonverbal modalities in a balanced and strongly interwoven manner.

4.2 A Dancer

In a recent application built at HMI, a virtual human - the Virtual Dancer - invites a real partner to dance with her [59]. The Virtual Dancer dances together with a human 'user', aligning its motion to the beat in the music input and responding to whatever the human user is doing. The system observes the movements of the human partner by using a dance pad to register feet activity and the computer vision system to gain information about arm and body movements. Using several robust processors, the system extracts global characteristics about the movements of the human dancer like how much (s)he moves around or how much (s)he waves with the arms. Such characteristics can then be used to select moves from the database that are in some way 'appropriate' to the dancing style of the human dancer.

There is a (non-deterministic) mapping from the characteristics of the observed dance moves to desirable dance moves of the Virtual Dancer. The interaction model reflects the intelligence of the Virtual Dancer. By alternating patterns of following the user or taking the lead with new types of dance moves, the system attempts to achieve a mutual dancing interaction where both human and virtual dancer influence each other. Finding the appropriate nonverbal interaction patterns that allow us to have a system that establishes rapport with its visitors is one of the longer term issues being addressed in this research.

Clearly, the domain of dancing is interesting for animation technology. We however focus on the interaction between human and virtual dancer. The interaction needs to be engaging, that is, interesting and entertaining. Efficiency and correctness are not the right issues to focus on. In this interaction perfectness can become boring and demotivating. First experiences with demonstration setups at exhibitions indicate that people are certainly willing to react to the Virtual Dancer.

4.3 A Conductor

We have designed and implemented a virtual conductor [9] that is capable of leading, and reacting to, live musicians in real time. The conductor possesses knowledge of the music to be conducted, and it is able to translate this knowledge to gestures and to produce these gestures. The conductor extracts features from the music and reacts to them, based on information of the knowledge of the score. The reactions are tailored to elicit the desired response from the musicians.

Clearly, if an ensemble is playing too slow or too fast, a (human) conductor should lead them back to the correct tempo. She can choose to lead strictly or more leniently, but completely ignoring the musicians' tempo and conducting like a metronome set at the right tempo will not work. A conductor must incorporate some sense of the actual tempo at which the musicians play in her conducting, or else she will lose control. If the musicians play too slowly, the virtual conductor will conduct a little bit faster than they are playing. When the musicians follow him, she will conduct faster yet, till the correct tempo is reached again.

The input of the virtual conductor consists of the audio from the human musicians. From this input volume and tempo are detected. These features are

evaluated against the original score to determine the conducting style (lead, follow, dynamic indications, required corrective feedback to musicians, etc) and then the appropriate conducting movements of the virtual conductor are generated. A first informal evaluation showed that the Virtual Conductor is capable of leading musicians through tempo changes and of correcting tempo mistakes from the musicians. Computer vision has not been added to the system. That is, musicians can only interact with the conductor through their music. In ongoing work we are looking at issues such as the possibility to have the conducting behavior directed to (the location of) one or more particular instruments and their players, experimenting with different 'corrective conducting strategies' and extending the expression range of the Virtual Conductor.

4.4 A Trainer

The Virtual Trainer (VT) application framework is currently under development [67] and involves a virtual human on a PC, who presents physical exercises that are to be performed by a user, monitors the users performance, and provides feedback accordingly at different levels. Our VT should fulfill most of the functions of a real trainer: it not only demonstrates the exercises to be followed, it should also provide professionally and psychologically sound, human-like coaching. Depending on the motivation and the application context, the exercises may be general fitness exercises that improve the users physical condition, special exercises to be performed from time to time during work to prevent for example RSI (Repetitive Strain Injury), or physiotherapy exercises with medical indications. The focus is on the reactivity of the VT, manifested in natural language comments relating to readjusting the tempo, pointing out mistakes or rescheduling the exercises. When choosing how to react, the static and dynamic characteristics of the user and the objectives to be achieved are to be taken into account and evaluated with respect to biomechanical knowledge and psychological considerations of real experts. For example, if the user is just slowing down, the VT will urge him in a friendly way to keep up with the tempo, acknowledge with cheerful feedback good performance and engage in a small talk every now and then to keep the user motivated.

Related work on VTs can be found in, among others, [18], where a physiotherapist is described with similar functionality as ours, [13] with an interesting Tai Chi application, and [2], reporting about work on an aerobics trainer.

5 How Human Should Virtual Humans Really Be?

Traditionally, there are different qualities associated with machines (including computers) and humans. Machines, and machine made products are praised for being reliably identical and precise, irrespective of time and conditions of production, exhaustive, fast, deterministic in handling huge amounts of information or products, slavishly programmable... as opposed to humans being nonrepetitive and less predictable, error prone and nondeterministic in their 'functioning'.

On the other hand, when it comes to flexibility, adaptation, error recovery, engagement... humans are valued higher.

However, it has been suggested that implicit interaction, which is an important part of the interaction in an AmI environment, should be inspired in the patterns and behaviour found in Human-Human interaction [36]. And, depending on the application, humans already treat computers as social actors [56]. Humanlike appearance and behavior of the interface, in particular when the computer is 'hidden behind a virtual human', strengthen this acceptance of the computer as a social actor [23; 47] or even a persuasive social actor [22] that changes our attitudes and behavior. Virtual Humans, more than regular graphical user interfaces, invite natural human interaction behavior and therefore it is useful to be able to capture and understand this behavior in order to let it play its natural role in an interaction [51].

This raises the question of how human Virtual Humans should be. Should a VH interface exhibit 'socially inspired' behaviour? Should a VH exhibit also the imperfections and shortcomings so often present in human communication? Do we want a VH to be hesitant or even make errors, to be nondeterministic, or, just the opposite, are we eager to use superhuman qualities made possible by the machine? Do we need the fallible human qualities to increase the naturalness and believability of human looking software entities? Are there additional, practical values too of human imperfections? Is there a third choice, namely VHs which unify the functionally useful capabilities of machines and humans, and thus are not, in principle, mere replicas of real humans?

In this paper we focus on qualities of real humans that are characteristically present in everyday life, but are hardly covered by the efforts on attaining higher level cognitive model based behavior of VHs. We look at subtleties and 'imperfections' inherent in human-human communication, and investigate the function and merits of them. By subtleties, we mean the rich variations we can employ in verbal and nonverbal communication to convey our message and the many ways we can draw attention to our intentions to convey a message. By imperfections we mean phenomena which are considered to be incorrect, imperfect according to the normative rules of human-human communication. Both enrich the model of communication, the first one by taking into account more aspects and details (e.g. emotions and personality of the speaker), the second one by being more permissive about what is to be covered. We consider imperfections as those natural phenomena of everyday, improvised language usage which are not considered to be correct and thus, are not 'allowed' by some descriptive rules of proper language usage. For instance, restarting a sentence after an incomplete and maybe grammatically incorrect fraction is such an imperfection. We are aware though of the continuous change of norms for a language, pushed by actual usage. The space here does not allow dwelling on the relationship between intelligence and the capability of error recovery and robust and reactive behavior, in general. Here we present the issue from the point of view of communication with virtual humans and from the point of view of perception of VHs.

6 Virtual Humans as Individuals

"Who are you?" do people ask (usually as one of the first questions) from their VH interlocutor. The answer is a name, maybe extended with the services the VH can offer. In case of chat bots, a date of birth may be given, and the creator may be named as 'father', such as in the case of Cybelle[1]. Notably, the date of 'creation' makes sense in the fictional framework only. Moreover, any inquiry about further family members is not understood. The personal history is similarly shallow and inconsistent as of her hobbies: she has a favorite author, but cannot name any title by him. Deviations from this common solution can be found, when the VH is to stand for a real, though long-dead, person [7], and the very application is to introduce the reincarnated real person and his history to the user. The other extreme is feasible when the VH is in a role, such as that of a museum guide [39], where his refusal 'to talk about any personal matters' sounds to be a natural reaction. But in other applications, where it would be appropriate, we would never know about the family, schooling, living conditions, acquaintances and other experiences of the VH, neither about his favorite food or hobbies. One may argue that that is enough, or even preferred, to remain 'to the point' in well defined task oriented application like a weather reporter or trainer. However, even in such cases in real life some well placed reference to the expert's 'own life and identity' breaks the businesslike monotonicy of the service, and can contribute to create common ground and build up trust. B. Hayes-Roth endowed her Extempo characters with some own history as part of their 'anima' [32]. From the recent past, we recall a Dutch weather forecast TV reporter who added, when a certain never heard-of Polish town was mentioned as the coldest place in Europe, that this town is special for him as his father was born there. But he could have noted about some other aspects like special food or customs he experienced or knows of from that place. In case of a real fitness trainer's video, it is remarkable how the task related talk is interwoven with references to the presenter's personal experience on where she learnt the exercises, what she found difficult, etc. A VH could use his personal background to generate just some 'idle small talk' in addition to the task related conversation, or to relate it to the stage of task completion or difficulty and the reactions from the user, in order to increase the user's commitment. So for instance, a VT may include non task related small talk at the beginning or during resting times, or add task related background information to keep the user motivated during a long and/or difficult exercise.

In order to make a VH 'personal', it is not enough to endow him with a 'personal history'. Some mechanisms should be provided to be able to decide when and what piece of personal information to tell. E.g. to derive if there is something in the personal knowledge of the VH which could be related to the factual, task oriented information to be told. This may span from simple tasks as discovering dates, names and locations, to the really complex AI task of associative and analogical reasoning.

[1] http://www.agentland.com/

Finally, the disclosure of the personal information and identity is a manifestation of personality: open, extrovert people (and VHs) may interweave more their story with personal references than introvert ones.

An interesting question is that a VH's 'personal history' may be also adapted to a situation, or a given user (group), not only its conversational style as suggested for robots [17] and VHs [63]. However, consistency within different interaction sessions with the same user (group) should be taken care of.

7 Here and Today - Situatedness

VHs hardly give the impression that they know about the time and situation they converse in with their user. Some VHs do reflect the time of the day by choosing an appropriate greeting. But much more could be done: keeping track of the day, including holidays, and commenting accordingly, providing 'geographical update' capability when placing a VH enabled service in a location, endowed with some social and political information about the place. Imagine a VT who knows that it is today a public holiday in Italy where the given VT is 'active'. Some special words to the user keeping up her exercise scheme on a holiday would be appropriate. But on a tropical Summer day, the heat may lead the VT to revise its strategy, remind the user the necessity of drinking, or even shorten the exercises, or suggest doing it the next time during the morning.

The identity of the user may be a source of further situatedness. As a minimum, a VH should 'remember' earlier encounters with the user. Asking the name or telling the same piece of small talk to the same person each time is disappointing. But how nice it sounds if a VT refers to yesterday's performance, knows of the user's religion does not allowing her to do exercises on Saturday, greets her specially on her birthday.

Finally, in order to perceive a VH as 'present', the VH must have means to gather information about the user and react to it. To begin with, the mere presence of the user and her identity should be detected, and her task related performance should be monitored. But think of a real trainer or tutor, who would very likely comment on changes like not wearing glasses, change in hair style, being sunburnt or showing signs of a cold. A Virtual Trainer could do similar comments.

8 Shortcomings as Features

'Shortcoming' is a negative word. Shortcomings in the realisation of communicative acts (e.g. mumbling, sloppy gestures ...), the information transfer (e.g. ambiguous wording, redundant or incomplete content ...) or the knowledge or decisions: at first sight they are something to be avoided when you are designing VHs. But... people without imperfections don't exist - luckily so, as the variety in 'imperfections' also makes people different and interesting. And besides, what is a shortcoming? Traits that are considered undesirable in one culture,

may be unimportant or even considered good behaviour in the next. And behavior that may be considered imperfect because it deviates from a norm imposed by 'average human behavior' may actually have a clear communicative function.

In Section 5, we raised the question whether we should aim for 'perfect' VHs or whether there are practical values to human imperfections too. In this section we want to elaborate a bit on the question, examining some examples of 'imperfect' behavior on their potential merits.

8.1 User Adaptation

At first sight one might require that a VH is as perfect as can be, in its cognitive decisions, conversational and motion skills. One reason for lifting this requirement may be user adaptation. As a VH may come across users of different intellectual and communicational capabilities, the capability to recognize such a situation and scale down the VHs' own functioning is more beneficial than overwhelming the user with a perfect and too demanding performance. This is already happening in the case of some TTS services, where the generated English pronunciation is not the 'official nicest' one, but one 'tortured' according to the practice of users whose mother tongue, like many Asiatic languages, has a very different acoustic scheme. Adapting to the level of a user by hiding some of the cognitive capabilities of the VH is already a usual practice in gaming and in tutoring situations. Lastly, the VT scenario exemplifies a possible reason for adapting the motion skills to those of the user by showing up in the embodiment the most appropriate for the given user. His/her gender, age and motion characteristics may be similar to the user's, in order to avoid having a too huge gap between or example a 'fit and young' trainer and the 'fattish, somewhat rigidmoving' user [71]. On the other hand, deviations from usual trainers in the other (superman) direction may have positive effects too. Imagine the VT making every now and then extreme, beyond-realistic capabilities jumps to cheer up the user or grab attention.

8.2 Clarification and Commitment

Certain types of imperfections in communication need 'repair'. A (virtual) human who mumbles needs to be asked to repeat itself. A (virtual) human using ambiguous language, or serious disfluencies, may be asked for clarification, literally, or through nonverbal mechanisms such as a lifted eyebrow or a puzzled expression. The traditional judgment of such repair in Human-Computer interaction is as an undesirable necessary evil, hence the term 'repair'. However, as Klein *et al.* state, one of the main criteria for entering into and maintaining a successful joint activity is "Commitment to an intention to generate a multiparty product" [37]. This commitment needs not only to be present, but must also be communicated to the conversational partner. Imperfections in the communications of a VH, and the subsequent so-called 'repair dialogues', could be a

subtle means for both VH and human to signal their commitment to the interaction. One must be really committed to an interaction if one is going through the trouble of requesting and/or giving clarifications, repeating oneself, etc...

There are two sides to this commitment issue. The first relates to imperfections at the side of the human user. When we assume that entering into clarification and repair dialogues is a strong signal of commitment to the conversation, we see a clear reason to invest in the development of techniques for clarification dialogues beyond what is needed to let the conversation reach its intended goal. One may decide to make the VH initiate clarification dialogues when they are not absolutely necessary for reaching the goal, to signal commitment to understanding the user.

The second relates to imperfections at the side of the VH. If the human is forced to ask the VH for clarification all the time, this will be considered a drawback in the interaction capabilities of the system. However, there may be certain subtle advantages to a VH that uses ambiguous, disfluent or otherwise imperfect expressions. Judicious use of ambiguities at non-critical points in the conversation, at a point where it is likely that the user will ask for clarification (explicitly or nonverbally), gives the VH a chance to show off its flexibility and its willingness to adapt to the user. This gives the human user a feeling of 'being understood' and of commitment from the side of the VH. Again, such an approach would need a lot of investment in repair and clarification capabilities of the VH.

8.3 Signaling Mental State and Attitude

In other situations imperfections express an important part of the content of the conversation. Imperfections in the multimodal generation process, such as hesitation, stuttering, mumbling and disfluencies can signal indications of the VH's cognitive state (e.g. 'currently thinking', 'unhappy'), conversational state ('ready to talk'), attitude towards the conversation partner or the content (belief, certainty, relevance, being apologetic).

Cognitive or conversational state is often reflected by gaze and body postures of the VHs. However, the usage of (non-speech) vocal elements in these and other situations has not been addressed widely yet. For example, by analyzing a multi-party real-life conversation, we found that non-speech elements abundantly interwove with the 'meaningful, articulated' utterances.

While some of the erroneous and nonverbal utterances reflect the 'processing deficiencies' of the speaker (e.g. difficulty in formulating a statement in correct format), others have important function in regulating the dialogue (e.g. indicating request for turn taking by making some sound) or in qualifying the 'verbatim' content (e.g. a hesitant pause before making a statement indicates that the information to be conveyed may not be correct). In our analysis of multi-party real-life conversation we could identify speech situations as well as personality, emotional and cognitive state of the speaker as indicators of the frequency and type of non-speech elements used.

8.4 Other Merits of Imperfections

The above subsections presented a number of considerations for turning VHs 'imperfect'. Before ending this section we will touch lightly upon a few of the countless other themes that in some way also can involve imperfections.

Disfluencies and other imperfections are a major characteristic of spontaneous speech. If a VH talks without disfluencies it may come across as stilted, not spontaneous enough. In the ongoing work mentioned earlier we do encounter a common phenomenon in the language usage itself: speakers do not express themselves in perfect sentences, especially if they are answering an unexpected, unusual question or contribute to a discussion. Often, they abandon an erroneous start of a sentence and correct, or repeat the start in another form.

Ambiguous wordings and underspecification will help to give the user a sense of 'freedom' in the dialogue, to feel less constrained [24]. (Note though that we then run the risk of the user indulging in the same kind of ambiguous language use to an extent that the VH cannot handle it). Hesitations, mumbling and disfluencies are also mechanisms that play an important role in reducing the amount of threat potentially perceived by the human user [10].

Making mistakes may have the positive side-effects of users perceiving the VH as more believable and more individual. Humans are not perfect, and usually there is not a single 'perfect' way of action. When researchers create a model of, for example, turn taking behavior as a general norm, and we then implement this model in a VH, the fact that such a model is an abstraction by nature makes certain that the VH will behave in a way that no real human would do. Making mistakes also gives a VH unrivalled opportunity for exhibiting an individual style. After all, as humans we can probably distinguish ourselves from others through the mistakes that we make as much as through the things we do perfect.

9 Some Communicational Challenges for VHs in an AmI Environment

Finally, there are some communicational challenges for Virtual Humans we would like to mention here. They have a lot to do with the fact that VHs in an AmI environment need to coordinate their actions carefully to events happening in the environment, as well as to the actions of the user (see also the paper by Jeffrey Cohn in this volume [16]). The last is especially tricky when the interaction is implicit, since then the user is not aware of the interaction and one cannot rely on the simple turntaking protocols that usually regulate interaction between a VH and a user.

9.1 Who Is in Control?

VHs enter domains where they are not to be the dominant partner, and thus the control scenario is not well established. On one hand, the VH should not be completely dominant and thus needs to be able to follow the initiative of the human partner. On the other hand, VHs that show no initiative of their own

do not seem to be human-like conversational partners but more like responsive machines that are to be controlled by the user using multimodal commands.

In the Virtual Dancer application, the control between human and VH is both mixed-initiative and implicit: the VH alternates phases of 'following dancing behavior' where it incorporates elements of the humans dance in its own dance with phases of 'leading dancing behavior' where it introduces new elements which the human will hopefully pick up.

For the Conductor, control is on the side of the VH, but with a special twist. Although the conductor is leading the musicians all the time, she cannot afford not to let herself be influenced by the actions of the musicians. If an ensemble is playing too slow or too fast, a conductor should lead them back to the correct tempo. She can choose to lead strictly or more leniently, but completely ignoring the musicians tempo and conducting like a metronome set at the right tempo will not work. A conductor must incorporate some sense of the actual tempo at which the musicians play in her conducting, or else she will lose control.

In the VT scenario, the issue of control is essential, and subtle. Basically, the user is to perform the instructions given by the VT. However, this is not what happens all the time. The reaction for the VT depends on the assessment of the situation, including past performance and knowledge of the user. When the VT concludes that the user has just lost tempo, or is getting a little lazy, than the VT reinforces his/her own tempo to the user. But if the user looks very exhausted, or (s)he has a 'bad day' of decreased performance, then the VT may 'give in' to the user and slow down his/her tempo to comfort the user. Hence the VT is attentive and reactive, instead of imposing a predefined scenario on the user. The decision concerning when and how 'to give in' to the user must be based on detailed domain-specific and psychological knowledge.

9.2 Coordination of Modalities

There are several reasons why coordination of modalities in our applications needs a level of subtlety and sophistication beyond what is present in much current work.

For example, in present VH applications it is usually assumed that speech is the leading modality, and accompanying gestures and facial expressions should be timed accordingly. Dropping this assumption influences the design of a balanced multi-modal behavior that does not place natural language in a privileged role. It also influences the planning algorithms used to generate the behavior. We know of only one work where the instruction utterances' tempo is scaled to the duration of the hand movements in explaining an assembly task [40]. This clearly indicates how rarely this issue is attended by current research efforts.

Another challenging characteristic is the need for alignment of multimodal behavior to external channels or events. Such a feature is essential in any application where the VH is embedded in an external environment or does not converse by switching between pre-programmed active speaking and passive listening behavior but reacts according to a model of bidirectional communication behavior. 'Outside events' are usually not covered in the framework of modality

coordination, but rather as an issue of planning specific gestures (e.g. pointing at or reaching for moving objects). However, the fact that behavior in any of the modalities may be constrained in its timing by sources outside the VHs' influence calls for subtle and strongly adaptive planning and re-planning.

In the Virtual Dancer application the issue of coordination focusses wholly on the alignment of the dance behavior to the music. There is also a relation between the performed dance 'moves' and the dance of the user, but as yet, there is no tight timing relation between those two.

Coordination in the Conductor application is really a mutual coordination between the intentions and actions of conductor on the one hand, and the musicians playing their parts on the other hand. One cannot simply say that the conductor coordinates her actions to the music as it is being played, nor is it realistic to assume that the conducting actions are planned first and the coordination of music played by the musicians is completely determined by that. There is a feedback loop of mutual influence going on: conducting and playing music is a joint action [15] in the truest sense.

In the VT scenario, the issue of coordinating speech, motion and a given piece of music has turned out to be central, relating to two aspects mentioned above. Namely, a virtual trainer explains postures and basic movements and conducts rhythmic exercises. The exercises may be performed in different tempi. The speech (e.g. counting and providing comments on posture), which is often of secondary importance, should be aligned to the motion. The alignment should be subtle, in one case making sure that the emphasized syllabe of the counting coincides with the end of the stroke of the movement, while in another case the counting starts together with a repetitive motion unit, or an expression is uttered in an elongated duration, during the finishing movement of a series. This requires real-time reactive planning of speech, occasionally resulting in what would usually be seen as 'unnaturally slow' utterances which are justified by the elongated duration of the corresponding motion. TTS systems may not even be prepared to generate such slowed down, unnatural speech. On the other hand, the timing of the exercise presented by the VT may be driven by external audiovisual cues of varying tempo. An exercise author may specify the tempo of an exercise by clapping or tapping, or as being aligned to music beats. However, this is often not enough in the case of VTs. For example, a VT may need to align its performance to the user doing the same exercise while counting along. This again requires a subtle real-time planning, affecting duration and alignment of sub-segments of a motion [66].

10 Discussion

We have argued that there are dimensions still to be exploited to turn VHs more lifelike, entertaining and in cases effective, especially in the context of Human Computing. We discussed the potentials of:

– making VHs look more individual and imperfect, may be configured to a given user's preferences;

– endowing VHs with identity and personal history;
– grounding VHs to the geographical and sociological place and time of the application being used;
– taking care of styled and natural conversation with phenomena of 'imperfections' reminiscent in real life.

The above features do not require, first of all, further improvement of single or multimodal communication on the signal level, but they do pose challenges on modeling mental capabilities like associative storytelling or require further socio-psychological studies of the nature and effect of social conversation in task related situations. What is needed, for several of the above enrichments, is multisignal perception of the conversant of a VH.

Dedicated evaluation studies are needed to put together a huge jigsaw image. It is clear already that the objective to engage the user in an activity and to perform a task well and efficiently may require a different VH design along several dimensions. Also, the application context (real/fictional) puts the user in different frame of mind to judge the VH, On the other hand, even less is known of the judgments of nonhuman capabilities of VHs. For example, it has turned out that a VH could 'read from the eye' of the user better than most of the people are capable of [49]. What to do with such a superhuman power of a VH? Or, another example is the reasoning speed and capability of a VH: do people take it as natural (from a VH) that he can recall multiple telephone books? Or should he 'fake' the human limitation of recalling data in a register? How to get away with shallow, or not deep/complete enough, models?

We have argued that carefully 'designed' imperfection in the communication may increase not only the believability but also the scope of applicability and effectiveness of the VH in question. People are are neither uniform nor perfect, but have different capabilities, and have means to recognize and cope with errors and limitations. Endowing VHs with the imperfections of humans can help making them more 'comfortable' to interact with. The natural communication of a VH should not be restricted to multimodal utterances that are always perfect, both in the sense of form and of content.

This scenario aims at hiding the 'machine' nature of the VH. This may be very much what we want on the communication level, but how about the cognitive level? For instance, if a VH is to find an item from a huge database, or perform a difficult calculation, should the imperfect that is, slower, error prone human behavior be mocked up, even if the computer is ready with the perfect answer immediately? In general, should a VH's amount and processing of knowledge 'resemble' the capabilities of humans, in order to make the VH believable and lifelike, as opposed to some omnipotent supercreature?

On the other hand, such a scenario may sound completely irrational, as it makes no use of the power of the computer. Given the fact that most humans are indoctrinated from birth with the adagio that 'computers are fast in calculating', hiding this capability behind artificial imperfection might even be perceived as unrealistic. Except maybe when the user is given to understand that (s)he

interacts with the computer through the intermediation of the VH rather than with the computer embodied by the VH.

Acknowledgments. We would like to thank the anonymous reviewers of earlier versions of this paper for their suggestions and discussions. This work is supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction, publication AMIDA-1). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] Abaci, T., de Bondeli, R., Cíger, J., Clavien, M., Erol, F., Gutiérrez, M., Noverraz, S., Renault, O., Vexo, F., and Thalmann, D. Magic wand and the enigma of the sphinx. *Computers & Graphics 28*, 4 (2004), 477–484.

[2] Babu, S., Zanbaka, C., Jackson, J., Chung, T., Lok, B., Shin, M. C., and Hodges, L. F. Virtual human physiotherapist framework for personalized training and rehabilitation. In *Proc. Graphics Interface 2005* (May 2005).

[3] Badler, N. LiveActor: A virtual training environment with reactive embodied agents. In *Proc. of the Workshop on Intelligent Human Augmentation and Virtual Environments* (October 2002).

[4] Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J. J., and Turk, M. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators and Virtual Environments 13*, 4 (2004), 428–441.

[5] Bailenson, J. N., Yee, N., Patel, K., and Beall, A. C. Detecting digital chameleons. *Computers in Human Behavior* (2007). in press.

[6] Baylor, A. L. S., and Ebbers, S. The pedagogical agent split-persona effect: When two agents are better than one. In *Proc of the World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA)* (2003).

[7] Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M. First prototype of conversational h.c. andersen. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces* (New York, NY, USA, 2004), ACM Press, pp. 458–461.

[8] Bickmore, T., and Cassell, J. Small talk and conversational storytelling in embodied interface agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence* (2000), pp. 87–92.

[9] Bos, P., Reidsma, D., Ruttkay, Z. M., and Nijholt, A. Interacting with a virtual conductor. In Harper et al. [29], pp. 25–30.

[10] Brown, P., and Levinson, S. C. *Politeness : Some Universals in Language Usage (Studies in Interactional Sociolinguistics)*. Studies in Interactional Sociolinguistics. Cambridge University Press, February 1987.

[11] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. F., Eds. *Embodied conversational agents*. MIT Press, Cambridge, MA, USA, 2000.

[12] Chafai, N. E., Pelachaud, C., Pelé, D., and Breton, G. Gesture expressivity modulations in an eca application. In Gratch et al. [27], pp. 181–192.

[13] Chao, S. P., Chiu, C. Y., Yang, S. N., and Lin, T. G. Tai chi synthesizer: a motion synthesis framework based on keypostures and motion instructions. *Computer Animation and Virtual Worlds 15*, 3-4 (2004), 259–268.

[14] Chi, D., Costa, M., Zhao, L., and Badler, N. The emote model for effort and shape. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000).

[15] Clark, H. H. *Using Language*. Cambridge University Press, 1996.

[16] Cohn, J. Foundations of human centered computing: Facial expression and emotion. In Huang et al. [33], pp. 5–12.

[17] Dautenhahn, K. *Socially Intelligent Agents in Human Primate Culture*. In Payr and Trappl [52], 2004, ch. 3, pp. 45–71.

[18] Davis, J. W., and Bobick, A. F. Virtual PAT: A virtual personal aerobics trainer. In *Proceedings of Workshop on Perceptual User Interfaces (PUI'98)* (New York, 1998), IEEE, pp. 13–18.

[19] de Rosis, F., Cavalluzzi, A., Mazzotta, I., and Novelli, N. Can embodied conversational agents induce empathy in users? In *Proc. of AISB'05 Virtual Social Characters Symposium* (April 2005).

[20] Egges, A., Molet, T., and Magnenat-Thalmann, N. Personalised real-time idle motion synthesis. In *PG '04: Proceedings of the Computer Graphics and Applications, 12th Pacific Conference on (PG'04)* (Los Alamitos, CA, USA, 2004), IEEE Computer Society, pp. 121–130.

[21] Ekman, P. *The argument and evidence about universals in facial expressions of emotion*. John Wiley, Chicelster, 1989, pp. 143–146.

[22] Fogg, B. J. *Persuasive Technology: Using Computers to Change What We Think and Do*. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, January 2003.

[23] Friedman, B., Ed. *Human Values and the Design of Computer Technology*. No. 72 in CSLI Publication Lecture Notes. Cambridge University Press, 1997.

[24] Gaver, W. W., Beaver, J., and Benford, S. Ambiguity as a resource for design. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2003), ACM Press, pp. 233–240.

[25] Gratch, J., and Marsella, S. Tears and fears: modeling emotions and emotional behaviors in synthetic agents. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents* (New York, NY, USA, 2001), ACM Press, pp. 278–285.

[26] Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., and Badler, N. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems 17*, 4 (2002), 54–63.

[27] Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., Eds. *Intelligent Virtual Agents, 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006, Proceedings* (2006), vol. 4133 of *Lecture Notes in Computer Science*, Springer.

[28] Gustafson, J., Bell, L., Boye, J., Lindström, A., and Wiren, M. The NICE Fairy-tale Game System. In *Proceedings of SIGdial 04* (April 2004).

[29] Harper, R., Rauterberg, M., and Combetto, M., Eds. *Proc. of 5th International Conference on Entertainment Computing, Cambridge, UK* (September 2006), no. 4161 in Lecture Notes in Computer Science, Springer Verlag.

[30] Hartmann, B., Mancini, M., Buisine, S., and Pelachaud, C. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *AAMAS* (2005), F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh, and M. Wooldridge, Eds., ACM, pp. 1095–1096.

[31] Hartmann, B., Mancini, M., and Pelachaud, C. Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture Workshop* (2005), S. Gibet, N. Courty, and J.-F. Kamp, Eds., vol. 3881 of *Lecture Notes in Computer Science*, Springer, pp. 188–199.

[32] Hayes-Roth, B., and Doyle, P. Animate characters. *Autonomous Agents and Multi-Agent Systems 1*, 2 (1998), 195–230.

[33] Huang, T., Nijholt, A., Pantic, M., and Pentland, A., Eds. *Proc. of the IJCAI Workshop on Human Computing, AI4HC07* (January 2007).

[34] Isbister, K., and Doyle, P. *The Blind Man and the Elephant Revisited: A Multidisciplinary Approach to Evaluating Conversational Agents.* Vol. 7 of Ruttkay and Pelachaud [62], 2004, ch. 1, pp. 3–26.

[35] Isla, D. A., and Blumberg, B. M. Object persistence for synthetic creatures. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems* (New York, NY, USA, 2002), ACM Press, pp. 1356–1363.

[36] Ju, W., and Leifer, L. The design of implicit interactions - making interactive objects less obnoxious. *Design Issues* (Draft). Draft for Special Issue on Design Research in Interaction Design.

[37] Klein, G., Feltovich, P. J., Bradshaw, J. M., and Woods, D. D. *Common Ground and Coordination in Joint Activity.* Wiley Series in Systems Engineering and Management. John Wiley and sons, Hoboken, New Jersey, 2004, ch. 6, pp. 139–178.

[38] Koda, T., and Maes, P. Agents with faces: the effect of personification. In *5th IEEE International Workshop on Robot and Human Communication* (November 1996), pp. 189=–194.

[39] Kopp, S., Jung, B., Lessmann, N., and Wachsmuth, I. Max - a multimodal assistant in virtual reality construction. *KI 17*, 4 (2003), 11.

[40] Kopp, S., and Wachsmuth, I. Model-based animation of coverbal gesture. In *CA '02: Proceedings of the Computer Animation* (Washington, DC, USA, 2002), IEEE Computer Society, p. 252.

[41] Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. The persona effect: affective impact of animated pedagogical agents. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1997), ACM Press, pp. 359–366.

[42] Lim, M. Y., Aylett, R., and Jones, C. M. Emergent affective and personality model. In *IVA* (2005), T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, Eds., no. 3661 in Lecture Notes in Computer Science, Springer, pp. 371–380.

[43] Mateas, M., and Stern, A. Façade: An experiment in building a fully-realized interactive drama, 2003.

[44] McBreen, H., Shade, P., Jack, M. A., and Wyard, P. J. Experimental assessment of the effectiveness of synthetic personae for multi-modal e-retail applications. In *Proceedings of the fourth international conference on Autonomous agents* (2000), pp. 39–45.

[45] Moppes, V. V. Improving the quality of synthesized speech through mark-up of input text with emotions. Master's thesis, VU, Amsterdam, 2002.

[46] Moundridou, M., and Virvou, M. Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning 18*, 3 (2002), 253–261.

[47] Nass, C., Isbister, K., and Lee, E.-J. *Truth is beauty: researching embodied conversational agents.* In Cassell et al. [11], 2000, ch. 13, pp. 374–402.

[48] Nijholt, A. Where computers disappear, virtual humans appear. *Computers & Graphics 28*, 4 (2004), 467–476.

[49] Nischt, M., Prendinger, H., André, E., and Ishizuka, M. Mpml3d: A reactive framework for the multimodal presentation markup language. In Gratch et al. [27], pp. 218–229.

[50] Noot, H., and Ruttkay, Z. M. Gesture in style. In *Gesture-Based Communication in Human-Computer Interaction* (2004), A. Camurri and G. Volpe, Eds., vol. 2915 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 324–337.

[51] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. *Human Computing and Machine Understanding of Human Behaviour: A Survey*, vol. 4451 of *Lecture Notes on Artificial Intelligence, Spec. Vol. AI for Human Computing*. 2007.

[52] Payr, S., and Trappl, R., Eds. *Agent Culture. Human-Agent Interaction in a Multicultural World*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2004.

[53] Plantec, P. M., and Kurzweil, R. *Virtual Humans*. AMACOM/American Management Association, November 2003.

[54] Poggi, I., Pelachaud, C., and De Carolis, B. To display or not to display? towards the architecture of a reflexive agent. In *Proceedings of the 2nd Workshop on Attitude, Personality and Emotions in User-adapted Interaction. User Modeling 2001* (July 2001), pp. 13–17.

[55] Prendinger, H., and Ishizuka, M. Social role awareness in animated agents. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents* (New York, NY, USA, 2001), ACM Press, pp. 270–277.

[56] Reeves, B., and Nass, C. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA, 1996.

[57] Rehm, M., and André, E. Catch me if you can: exploring lying agents in social settings. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems* (New York, NY, USA, 2005), ACM Press, pp. 937–944.

[58] Reidsma, D., op den Akker, H. J. A., Rienks, R., Poppe, R., Nijholt, A., Heylen, D., and Zwiers, J. Virtual meeting rooms: From observation to simulation. *AI & Society, The Journal of Human-Centred Systems to appear* (June 2007).

[59] Reidsma, D., van Welbergen, H., Poppe, R., Bos, P., and Nijholt, A. Towards bi-directional dancing interaction. In Harper et al. [29], pp. 1–12.

[60] Rienks, R., Nijholt, A., and Reidsma, D. Meetings and meeting support in ambient intelligence. In *Ambient Intelligence, Wireless Networking and Ubiquitous Computing*, T. A. Vasilakos and W. Pedrycz, Eds., Mobile communication series. Artech House, Norwood, MA, USA, 2006, ch. 17, pp. 359–378.

[61] Ruttkay, Z. M., and Noot, H. Animated chartoon faces. In *NPAR '00: Proceedings of the 1st international symposium on Non-photorealistic animation and rendering* (New York, NY, USA, 2000), ACM Press, pp. 91–100.

[62] Ruttkay, Z. M., and Pelachaud, C., Eds. *From Brows to Trust: - Evaluating Embodied Conversational Agents,*, vol. 7 of *Kluwers Human-Computer Interaction Series*. Kluwer Academic Publishers, Dordrecht, 2004.

[63] Ruttkay, Z. M., Pelachaud, C., Poggi, I., and Noot, H. *Excercises of Syle for Virtual Humans*. Advances in Consciousness Research Series. John Benjamins Publishing Company, to appear.

[64] Ruttkay, Z. M., Reidsma, D., and Nijholt, A. Human computing, virtual humans and artificial imperfection. In *ACM SIGCHI Proc. of the ICMI Workshop on Human Computing* (New York, USA, November 2006), F. Quek and Y. Yang, Eds., ACM, pp. 179–184.

[65] Ruttkay, Z. M., Reidsma, D., and Nijholt, A. Unexploited dimensions of virtual humans. In Huang et al. [33], pp. 62–69.

[66] Ruttkay, Z. M., and van Welbergen, H. On the timing of gestures of a virtual physiotherapist. In *Proc. of the 3rd Central European Multimedia and Virtual Reality Conference* (November 2006), Pannonian University Press, pp. 219–224.

[67] Ruttkay, Z. M., Zwiers, J., van Welbergen, H., and Reidsma, D. Towards a reactive virtual trainer. In *Proc. of the 6th International Conference on Intelligent Virtual Agents, IVA 2006* (Marina del Rey, CA, USA, 2006), vol. 4133 of *LNAI*, Springer, pp. 292–303.

[68] Si, M., Marsella, S., and Pynadath, D. V. Thespian: Modeling socially normative behavior in a decision-theoretic framework. In Gratch et al. [27], pp. 369–382.

[69] Stronks, B., Nijholt, A., van der Vet, P. E., and Heylen, D. Designing for friendship: Becoming friends with your eca. In *Proc. Embodied conversational agents - let's specify and evaluate them!* (Bologna, Italy, 2002), A. Marriott, C. Pelachaud, T. Rist, Z. M. Ruttkay, and H. H. Vilhjálmsson, Eds., pp. 91–96.

[70] Thórisson, K. R. *Communicative humanoids: a computational model of psychosocial dialogue skills*. PhD thesis, MIT Media Laboratory, 1996.

[71] van Vugt, H. C., Konijn, E. A., Hoorn, J. F., and Veldhuis, J. Why fat interface characters are better e-health advisors. In Gratch et al. [27], pp. 1–13.

[72] Wachsmuth, I., and Kopp, S. Lifelike gesture synthesis and timing for conversational agents. In *Gesture Workshop* (2001), I. Wachsmuth and T. Sowa, Eds., vol. 2298 of *Lecture Notes in Computer Science*, Springer, pp. 120–133.

[73] Walker, J. H., Sproull, L., and Subramani, R. Using a human face in an interface. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1994), ACM Press, pp. 85–91.

[74] Wang, J., Drucker, S. M., Agrawala, M., and Cohen, M. F. The cartoon animation filter. *ACM Transactions on Graphics 25*, 3 (2006), 1169–1173.

Affect Detection and an Automated Improvisational AI Actor in E-Drama

Li Zhang[1], Marco Gillies [2], John A. Barnden[3], Robert J. Hendley[3], Mark G. Lee[3], and Alan M. Wallington[3]

[1] School of Computing and Technology, University of East London, Dockland Campus, 4-6 University Way, London, E16 4LZ
[2] Department of Computer Science, University College London, London, WC1E 6BT
[3] School of Computer Science, University of Birmingham, Birmingham, B15 2TT
L.Zhang@cs.bham.ac.uk

Abstract. Enabling machines to understand emotions and feelings of the human users in their natural language textual input during interaction is a challenging issue in Human Computing. Our work presented here has tried to make our contribution toward such machine automation. We report work on adding affect-detection to an existing e-drama program, a text-based software system for dramatic improvisation in simple virtual scenarios, for use primarily in learning contexts. The system allows a human director to monitor improvisations and make interventions, for instance in reaction to excessive, insufficient or inappropriate emotions in the characters' speeches. Within an endeavour to partially automate directors' functions, and to allow for automated affective bit-part characters, we have developed an affect-detection module. It is aimed at detecting affective aspects (concerning emotions, moods, value judgments, etc.) of human-controlled characters' textual "speeches". The work also accompanies basic research into how affect is conveyed linguistically. A distinctive feature of the project is a focus on the metaphorical ways in which affect is conveyed. Moreover, we have also introduced how the detected affective states activate the animation engine to produce gestures for human-controlled characters. The description of our approach in this paper is taken in part from our previous publications [1, 2] with new contributions mainly on metaphorical language processing (practically and theoretically), 3D emotional animation generation and user testing evaluation. Finally, Our work on affect detection in open-ended improvisational text contributes to the development of automatic understanding of human language and emotion. The generation of emotional believable animations based on detected affective states and the production of appropriate responses for the automated affective bit-part character based on the detection of affect contribute greatly to the ease and innovative user interface in e-drama, which leads to high-level user engagement and enjoyment.

Keywords: E-drama, affect detection, improvisational AI actor, emotional behaviour and metaphor.

1 Introduction

Improvised drama and role-play are widely used in education, counselling and conflict resolution. Researchers have explored frameworks for e-drama, in which

T.S. Huang et al. (Eds.): AI for Human Computing, LNAI 4451, pp. 339–358, 2007.

virtual characters (avatars) interact under the control of human actors. The springboard for our research is an existing system (*edrama*) created by one of our industrial partners, used in schools for creative writing and teaching in various subjects. The experience suggests that e-drama helps students lose their usual inhibitions, because of anonymity etc. One main aspect of our project is the addition of types of intelligent automation.

In the *edrama* system, up to five virtual characters are controlled on a virtual stage by human users ("actors"), with characters' (textual) "speeches" typed by the actors operating the characters. A director is also involved in a session. A graphical interface on each actor's and director's terminal shows the stage and characters. Speeches are shown as text bubbles. Actors choose their characters clothes and bodily appearance.

Currently, cartoon figures against backdrops of real-life photographic images are used. However, we are bringing in animated gesturing avatars and 3D computer-generated settings using technology from an industrial partner. Actors and the human director work through software clients connecting with the server. Clients communicate using XML stream messages via the server, which is usually remote from the terminals, which may themselves be remote from each other. Terminal-server communication is over the Internet using standard browsers.

The actors are given a loose scenario around which to improvise, but are at liberty to be creative. One scenario we have used is school-bullying where a schoolgirl Lisa is being bullied by her classmate Mayid. There are also roles for two friends and a teacher. Within these parameters, actors must improvise interesting interchanges.

The human director has a number of roles. S/he must constantly monitor the unfolding drama and the actors' interactions, or lack of them, in order to check whether they are keeping to the general spirit of the scenario. If this is not happening, the director may then intervene. For example, a director may intervene when the emotions expressed or discussed by characters are not as expected (or are not leading consistently in a new interesting direction). The director may also feel the need to intervene if one character is not getting involved, or is dominating the improvisation.

Intervention can take a number of forms. The director can send messages to actors. However, another important means of directorial intervention is for the director to introduce and control a 'bit-part' character. This character will not have a major role in the drama, but might, for example, try to interact with a character who is not participating much in the drama or who is being ignored by the other characters. Alternatively, it might make comments intended to 'stir up' the emotions of those involved, or, by intervening, diffuse any inappropriate exchange developing.

Clearly, all this places a heavy burden on the director. In particular, playing the role of the bit-part character and interacting with other characters whilst keeping interventions limited so as to maintain the main improvisatory drama amongst the actors, makes it difficult to fully monitor the behaviour of all the other actors and send appropriate messages to them should they stray off topic or exhibit inappropriate emotions. The difficulty is particularly acute if the directors are novices, such as teachers trying to use e-drama in their lessons.

One major research aim is accordingly to automate some directorial functions, either to take some of the burden away from a human director, or to provide a fully automated (though necessarily very restricted) director. With a fully-automated director, even if highly restricted in what it could do, little or no human supervision

might be required for at least minimally adequate improvisations, and *edrama* could, for example, be added to websites about certain topics allowing visitors to engage in on-line role-play germane to the topic. However, our main current work is on merely assisting a human director. The assistance is by

(a) fully-automated control of an optionally-included bit-part character
(b) sending of automated suggestions to the human director about the progress of the improvisation or about messages to send to the human actors.

Point (b) is addressed briefly below. Our main focus in this paper is on (a).

For purpose (a), we have created a simple automated actor, EMMA, which controls a bit-part character who is an acquaintance of the other characters. EMMA contains an affect-detection module, which tries to identify affect in characters' speeches, allowing the EMMA character to make responses that, it is hoped, will stimulate the improvisation, thus leading to less need for intervention by the human director. Within affect we include: basic and complex *emotions* such as anger and embarrassment respectively; *meta-emotions* (emotions about emotions) such as desiring to overcome anxiety; *moods* such as hostility; and *value judgments* (judgments of goodness, importance, etc.). Although merely detecting affect is limited compared to extracting the full meaning of characters' utterances, we have found that in many cases this is sufficient for the purposes of stimulating the improvisation.

Even limited types of affect detection can be useful. We do not purport to be able to make EMMA detect all types of affect under all ways affect can be expressed or implied, or to do it with a high degree of reliability. The spirit of the project is to see how far we can get with practical processing techniques, while at the same time investigating theoretically the nature of, and potential computational ways of dealing with, forms of affective expression that may be too difficult currently to handle in a usable implemented system.

Much research has been done on creating affective virtual characters in interactive systems. Picard's work [3] makes great contributions to building affective virtual characters. Also, emotion theories, particularly that of Ortony et al. [4] (OCC), have been used widely therein. Ortony et al. include a rule-based system for the generation of the 22 emotion types, which is widely used for emotion generation for the development of intelligent virtual agents. In Prendinger and Ishizuka's work [5], animated agents are capable to perform affective communication. They have defined eliciting condition rules for more frequently used OCC emotion types. Then the generated emotional states are filtered by two control states – personality and social role awareness – in order to achieve believable emotional expression. Wiltschko's *eDrama Front Desk* [6] is an online emotional natural language dialogue simulator with a virtual reception interface for pedagogical purposes. Natural language parser is used to analyze users' input texts which index into a list of phrases that are frequently used. Then the system's dialogue manager selects an output phrase for the computer character. The emotion model of this system is derived from OCC model based on two factors: respect and power. Mehdi et al. [7] combined the widely accepted five-factor model of personality [8], mood and OCC in generating emotional behaviour for a fireman training application. Personality and mood have also played important roles in emotion generation and emotion intensity adaptation. Implementation of animated agents with emotional behaviour is their next step. Gratch and Marsella [9] presented

an integrated model of appraisal and coping, in order to reason about emotions and to provide emotional responses, facial expressions and potential social intelligence for virtual agents. Their main contribution is the introduction of computational description of emotional coping for the first time. They have also extended the scope of discourse on appraisal theories by incorporating influence between cognition and appraisal to obtain emotional coping. Additionally, Egges et al. [10] have also provided virtual characters with conversational emotional responsiveness. Elliott et al. [11] demonstrated tutoring systems that reason about users' emotions. Aylett et al. [12] also focused on the development of affective behaviour planning for the synthetic characters. There is much other work in a similar vein.

Additionally, human computing focuses on the creation of innovative interfaces to facilitate users to engage in more natural and inspiring interaction with machines by providing automatic understanding of human affect, language and behaviour [32]. Emotion recognition in speech and facial expression has been studied extensively [32, 33, 34]. But very few research work has made an attempt to dig out the affect flavour in human open-ended linguistic textual input in online role-play, although the first interaction system based on natural language textual input, Weizenbaum's Eliza [35], was first developed back in 1966. Thus there has been only a limited amount of work directly comparable to our own, especially given our concentration on improvisation and open-ended language. However, *Façade* [13] included shallow natural language processing for characters' open-ended utterances. There are two language processing components in *Façade*. The first component did many-to-few mapping and transformed various user text input into a few discourse acts. Then the second language processing component produced several reactions according to the generated discourse acts. Finally, the highest priority reaction in the highest priority context is chosen as the response for the virtual agents. But in *Façade*, the detection of major emotions, rudeness and value judgements is not mentioned. Zhe and Boucouvalas [14] demonstrated an emotion extraction module embedded in an Internet chatting environment (see also [15]). It uses a part-of-speech tagger and a syntactic chunker to detect the emotional words and to analyse emotion intensity for the first person (e.g. 'I' or 'we'). Unfortunately the emotion detection focuses only on emotional adjectives, and does not address deep issues such as figurative expression of emotion. Also, the concentration purely on first-person emotions is narrow. We might also mention work on general linguistic clues that could be used in practice for affect detection (e.g. [16]).

Our work is distinctive in several respects. Our interest is not just in (a) the positive first-person case: the affective states that a virtual character X implies that it has (or had or will have, etc.), but also in (b) affect that X implies it lacks, (c) affect that X implies that other characters have or lack, and (d) questions, commands, injunctions, etc. concerning affect. We aim also for the software to cope partially with the important case of metaphorical conveyance of affect [17, 18].

Our project does not involve using or developing deep, scientific models of how emotional states, etc., function in cognition. Instead, the deep questions investigated are on linguistic matters such as the metaphorical expression of affect. In studying how ordinary people understand and talk about affect in ordinary life, what is of prime importance is their *common-sense* views of how affect works, irrespective of scientific accuracy of those views. Metaphor is strongly involved in such views.

It should also be appreciated that this paper does not address the emotional, etc. states of the *actors* (or director, or any audience). Our focus is on the affect that the actors make their characters express or mention. While an actor may work him/herself up into, or be put into, a state similar to or affected by those in his/her own characters' speeches or those of other characters, such interesting effects, which go to the heart of the dramatic experience, are beyond the scope of this paper, and so is the possibility of using information one might be able to get about actors' own affective states as a hint about the affective states of their characters or vice-versa.

Various characterizations of emotion are used in emotion theories. The OCC model uses emotion labels (anger, etc.) and intensity, while Watson and Tellegen [19] use positivity and negativity of affect as the major dimensions. We have drawn ideas from several such sources. We use an evaluation dimension (negative-positive), affect labels, and intensity. The basic emotion labels (such as 'angry') we use are taken from Ekman [20], while other comparatively complex affect labels (such as 'approving') are taken from the OCC model. There are 25 affect labels used in our system currently. Affect labels plus intensity are used when strong text clues signalling affect are detected, while the evaluation dimension plus intensity is used when only weak text clues are detected. The description of our approach in this paper is taken in part from our previous publications [1, 2] with new contributions mainly on metaphorical figurative language processing (practically and theoretically), 3D animation generation and user testing evaluation.

Finally, machines have struggled to understand human language and emotion expressed in it. Expecting machines to engage in a meaningful role-play with human characters based on the affect understanding in drama improvisation is even unthinkable. Our work presented here made some efforts towards these goals. Although we have implemented a limited degree of affect-detection in an automated bit-part character in an e-drama application, the statistical analysis results in our user study, especially on user engagement and enjoyment, indicate that the improvisational AI actor, EMMA, performed as good as another secondary school student.

2 Our Current Affect Detection

Before any automatic recognition and response components could be built for use in our automated actor EMMA, a detailed analysis of the language used in edrama sessions was necessary. A small corpus of sessions was analysed by hand to identify the range of linguistic forms used and to provide insight for the automatic processing. In fact, this analysis is often very difficult and unreliable but it does reveal some important observations.

- The language used is often complex and idiosyncratic. It is almost invariably ungrammatical, it uses abbreviations, it contains mis-spellings and it borrows heavily from the language of text-messaging (*textese*) and chat-rooms. Compared to the language normally analysed in computational linguistics it provides significant additional challenges.
- The literal meaning of the statements is often less important to its interpretation than the affect that it is expressing. The content of a statement is still important to building an understanding and to responding appropriately, but understanding the affective state being expressed is critical.

- The language contains a large number of weak cues to the affect that is being expressed. These cues may be contradictory or they may work together to enable a stronger interpretation of the affective state. In order to build a reliable and robust analyser of affect it is necessary to undertake several diverse forms of analysis and to enable these to work together to build stronger interpretations.

This leads to a system where the emphasis is moved away from building a representation of the meaning of the statement to one where more weight is given to building a robust representation of affective connotations.

The results of this affective analysis are then used to:

- Control an automated actor (EMMA) that operates a character in the improvisation: see (a) in Section 1.
- Independently of this, help create the directorial suggestions mentioned in Section 1, in point (b).
- Additionally, drive the animations of the avatars in the user interface so that they react bodily in ways that is consistent with the affect that they are expressing, for instance by changing posture or facial expressions.

The overall architecture is shown in Fig. 1.

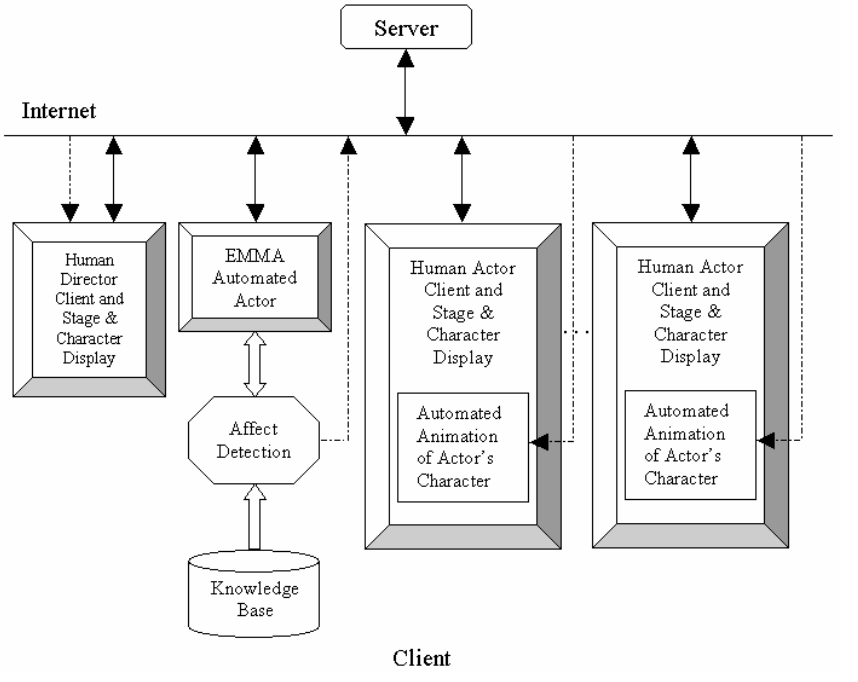


Fig. 1. Application architecture: Fat arrows show flow of information inside a program. Solid thin arrows show flow of character utterances and director messages formatted in XML. Dashed lines show flow of information about detected affect and flow of automated suggestions to the director.

Within the affect detection component we need to undertake several analyses of any given utterance by another character. These will each build representations which may be used by other components (e.g. syntactic structure) and will construct (possibly weak) hypotheses about the affective state. The architecture adopted is a blackboard based one. Each knowledge source undertakes its processing and writes its results to a central data structure (the blackboard) where they can be used by other knowledge sources and where hypotheses can be supported by multiple knowledge sources.

A rule-based component takes these hypotheses and builds a single interpretation of the affective state being expressed by the utterance being analysed. This interpretation is then transmitted to the other components of the system: EMMA (the automated actor), the automated directorial-suggestion generator and the animation component in the software client that handles the actor whose uttrerance is being processed.

The response generation component of EMMA uses this interpretation to build its behaviour driven mainly by its role in the improvisation and the affect expressed in the statement to which it is responding.

Fig. 2. illustrates the overall structure of the language processing.

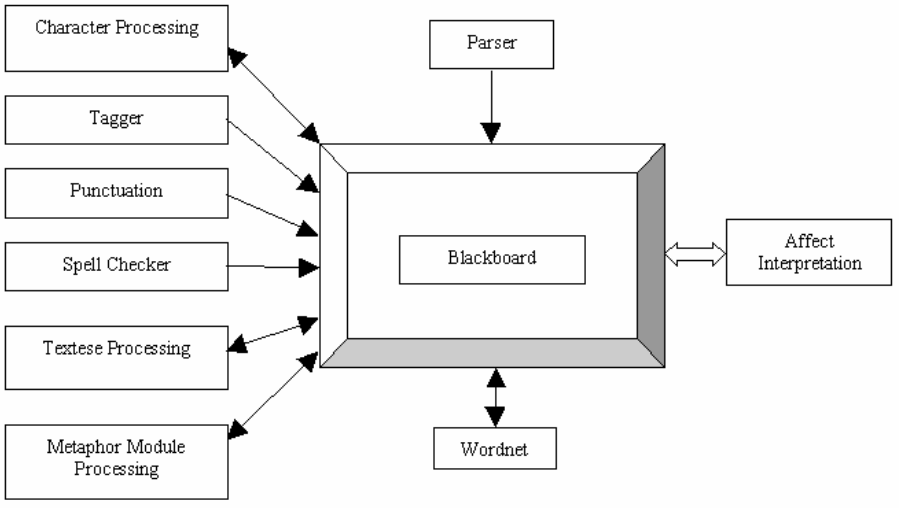


Fig. 2. Blackboard architecture

2.1 Pre-processing Modules

The language in the speeches created in *edrama* sessions severely challenges existing language-analysis tools if accurate semantic information is sought, even in the limited domain of restricted affect-detection. Aside from the complications noted above, the language includes slang, use of upper case and special punctuation (such as repeated exclamation marks) for affective emphasis, repetition of letters, syllables or words for emphasis, and open-ended interjective and onomatopoeic elements such as “hm”,

"ow" and "grrrr". In the examples we have studied, which so far involve teenage children improvising around topics such as school bullying, the genre is similar to Internet chat [21]. To deal with the misspellings, abbreviations, letter repetitions, interjections and onomatopoeia, several types of pre-processing occur before the main aspects of detection of affect. We have reported our work on pre-processing modules to deal with these language phenomena in detail in [1].

2.2 Affect Detection Using Rasp, Pattern Matching & WordNet and Responding Regimes

One useful pointer to affect is the use of imperative mood, especially when used without softeners such as 'please' or 'would you'. Strong emotions and/or rude attitudes are often expressed in this case. There are common imperative phrases we deal with explicitly, such as "shut up" and "mind your own business". They usually indicate strong negative emotions. But the phenomenon is more general.

Detecting imperatives accurately in general is by itself an example of the non-trivial problems we face. Expression of the imperative mood in English is surprisingly various and ambiguity-prone, as illustrated below. We have used the syntactic output from the *Rasp* parser [22] and semantic information in the form of the semantic profiles for the 1,000 most frequently used English words [23] to deal with certain types of imperatives. Briefly, the grammar of the 2002 version of the *Rasp* parser that we have used incorrectly recognised certain imperatives (such as "you shut up", "Dave bring me the menu" etc) as declaratives. We have made further analysis of the syntactic trees produced by *Rasp* by considering of the nature of the sentence subject, the form of the verb used, etc, in order to detect imperatives. We have also made an effort to deal with one special case of ambiguities: a subject + a verb (for which there is no difference at all between the base form and the past tense form) + "me" (e.g. 'Lisa hit/hurt me'.). The semantic information of the verb obtained by using Heise's [23] semantic profiles, the conversation logs and other indicators implying imperatives help to find out if the input is an imperative or not.

In an initial stage of our work, affect detection was based purely on textual pattern-matching rules that looked for simple grammatical patterns or templates partially involving specific words or sets of specific alternative words. This continues to be a core aspect of our system but we have now added robust parsing and some semantic analysis, including but going beyond the handling of imperatives discussed above.

A rule-based Java framework called Jess is used to implement the pattern/template-matching rules in EMMA allowing the system to cope with more general wording. In the textual pattern-matching, particular keywords, phrases and fragmented sentences are found, but also certain partial sentence structures are extracted. This procedure possesses the robustness and flexibility to accept many ungrammatical fragmented sentences and to deal with the varied positions of sought-after phraseology in characters' utterances. The rules conjecture the character's emotions, evaluation dimension (negative or positive), politeness (rude or polite) and what response EMMA should make. The rule sets created for one scenario have a useful degree of applicability to other scenarios, though there will be a few changes in the related knowledge database according to the nature of specific scenarios. The scenarios that we have used so far are school bullying and Crohn's disease.

However, it lacks other types of generality and can be fooled when the phrases are suitably embedded as subcomponents of other grammatical structures. In order to go beyond certain such limitations, sentence type information obtained from the Rasp parser has also been adopted in the pattern-matching rules. This information not only helps EMMA to detect affective states in the user's input (see the above discussion of imperatives), and to decide if the detected affective states should be counted (e.g. affects detected from conditional sentences won't be valued), but also helps EMMA to make appropriate responses.

Additionally, the sentence type information can also help to avoid the activation of multiple rules, which could lead to multiple detected affect results for one user's input. Mostly, it will help to activate only the most suitable rule to obtain the speaker's affective state and EMMA's response to the human character.

For example, as we discussed in the above, we use *Rasp* to indicate imperative sentences, such as in the school bullying scenario, when Mayid (*the bully*) said "Lisa, don't tell Miss about it". The pseudo-code example rule for such input is as follows:

```
(defrule example_rule
  (any string containing negation and the sentence type is 'imperative')
  =>
  (obtain affect and response from knowledge base))
```

Thus the declarative input such as "I won't tell Miss about it" won't be able to activate the example rule due to different sentence type information.

Additionally, a reasonably good indicator that an inner state is being described is the use of 'I' (see also [16]), especially in combination with the present or future tense (e.g. 'I'll scream', 'I hate/like you', and 'I need your help'). We especially process 'the first-person with a present-tense verb' statements using WordNet. When we fail to obtain the speaker's affective state in the current input by using Rasp and pattern matching, WordNet is used to find the synonyms of the original verb in the user's input. These synonyms are then refined by using Heise's [23] semantic profiles in order to obtain a subset of close synonyms. The newly composed sentences with the verbs in the subset respectively replacing the original verb, have extended the matching possibilities in the pattern-matching rules to obtain user's affective state in the current input.

After the automatic detection of users' affective states, EMMA needs to make responses in her role to the human characters during the improvisation. We have also created responding regimes for the EMMA character. Most importantly, EMMA can adjust its response likelihood according to how confident EMMA is about what it has discerned in the utterance at hand. Especially, in order to make contributions to the improvisation progression, EMMA also has a global view of the drama improvisation. Briefly, the knowledge base of EMMA (see Fig. 1.) provides scenario's background knowledge for each human character. EMMA can rise various scenario-related topics in its role for the human characters according to the detected affective states and topics discussed in the text input by using the rule-based reasoning based on the knowledge base. Inspection of the transcripts collected in the user testing indicates that EMMA usefully pushed the improvisation forward on various occasions (see section 4). For example, in the school bullying scenario, if the victim character hasn't shown the expected emotional states (such as fear or anger) in ones role-play (this

also means that this character lacks of expected negative emotional states), then the EMMA character will deliberately raise scenario-related topics towards this character in order to push the improvisation forward, e.g. by asking why he/she feels sad.

Details of the work reported in this section can be found in [1]. The brief summaries here of our previous implementations and their capabilities aim to remind readers.

2.3 Metaphorical Language Processing in EMMA

The metaphorical description of emotional states is common and has been extensively studied [17, 18]. E.g.: "He nearly exploded" and "Joy ran through me," where anger and joy are being viewed in vivid physical terms. Such examples describe emotional states in a relatively explicit if metaphorical way. But affect is also often conveyed more implicitly via metaphor, as in "His room is a cess-pit": affect (such as 'disgust') associated with a source item (cess-pit) gets carried over to the corresponding target item (the room). In other work, we have conducted research on metaphor in general (see, e.g. [24, 25]), and are now applying it to the e-drama application, and conversely using the application as a useful source of theoretical inspiration.

Our intended approach to metaphor handling in the EMMA affect-detection module is partly to look for stock metaphorical phraseology and straightforward variants of it, and partly to use a simple version of the more open-ended, reasoning-based techniques taken from the ATT-Meta project on metaphor processing [24]. As an example of stock phrase handling, insults in e-drama are often metaphorical, especially the case of animal insults ("you stupid cow", "you dog"). Particularly the 'second-person/a singular proper noun + present-tense copular form' statements (such as 'you're a rat', 'Lisa is a pig') and the second-person phrases (such as 'you dog') are often used to express insults. In the EMMA affect-detection module, we use *Rasp* to locate such user's input. We have also employed an on-line animal-name dictionary (http://dictionary.reference.com/writing/styleguide/animal.html), including names of animals, animal groups, young animals, etc, since usually calling someone a baby animal name (e.g. "puppy") may imply affection while calling someone an adult animal name could convey an insult. Then we use this animal-name dictionary to find out if there is any potential insulting/affectionate animal name present in the 'second-person/a singular proper noun + present-tense copular form' statements or in the second-person phrases. If there is, then we will use WordNet to analyze the animal name. If WordNet provides the semantic information of the animal name containing the description of the characteristics of a person/woman/men, such as 'an adjective + person/woman/man', then we will classify such user's input as metaphorical language. Additionally, we also use another semantic profile developed by Esuli and Sebastiani [26] to obtain the evaluation value of the adjective preceding the word 'person/woman/man'. If it's negative (e.g. 'a disagreeable person', 'a disgraceful person'), then we classify the user's input as metaphorical insulting language. If it's positive (e.g. 'a lovely person', 'a famous man'), then the user's input will be considered as metaphorical affectionate language. Otherwise, the user's input will be regarded as metaphorical objective language.

Not only do animal names present in the above statements may convey affective states, but also use of special person-types (e.g. 'freak') or mythical being (e.g.

'devil', 'angel') in such statements could also imply insults or approbations. Thus if there is a noun, but not an animal name, present in the 'second-person/a singular proper noun + present-tense copular form' statements or in the second-person phrases, then we collect all of its synonyms using WordNet. If any of the special person-types or mythical being collected in previous e-drama transcripts is shown among the synonyms, then we conclude that it contains insulting/affectionate flavour.

Sometimes, adjectives instead of nouns are used to directly convey affective states, such as 'Lisa is a stupid girl', 'you're a good mum' and 'you stupid boy'. If there is no noun present in the above statements or we couldn't obtain any semantic information by only analysing nouns in the above sentence structures, adjectives will be very helpful. We could find out the evaluation values of these adjectives, again using the semantic profile developed by Esuli and Sebastiani [26]. In this way, we may obtain at least the positive or negative flavour in the user's input.

One particular phenomenon of theoretical and practical interest is that physical size is often metaphorically used to emphasize evaluations, as in "you are a big bully", "you're a big idiot", and "you're just a little bully." The bigness is sometimes literal as well. "Big bully" expresses strong disapproval [27] and "little bully" can express contempt, although "little" can also convey sympathy or be used as an endearment. Such examples are not only important in practice but also theoretically challenging. We have also encountered surprisingly creative uses of metaphor in e-drama. For example, in the school-bullying scenario, Mayid is portrayed as having already insulted Lisa by calling her a 'pizza' (short for 'pizza-face'). This figurative insult was given a theoretically intriguing, creatively metaphorical elaboration in one improvisation, where Mayid said "I'll knock your topping off, Lisa."

Our work on metaphor outside the e-drama research is focused on an approach and system called ATT-Meta [24]. This approach is heavily dependent on detailed utterance-meaning analysis and on rich knowledge bases and reasoning processes, and is currently unsuitable for direct use in the *edrama* system. However, examples arising in *edrama* transcripts such as the more creative ones above provide useful data guiding the further development of ATT-Meta and can pose useful challenges to current metaphor theory generally. Moreover, theoretical study of the affective metaphorical language in e-drama transcripts is presented in section 3.

Finally, the detected affective states in the user's text input and EMMA's responses to other human characters have been encoded in an xml stream, which is sent to the server by EMMA. Then the server broadcasts the xml stream to all the clients so that the detected affective states information can be picked up by the animation engine to contribute to the production of 3D gestures and postures for the avatars. Now we will discuss the generation of emotional believable animation in detail in the following section.

2.4 The Users Avatars and Emotional Animation

The topics discussed in the edrama scenarios are often highly emotionally charged and this is reflected in the animation of the characters. Each participant in edrama has their own animated graphical character (avatar). In order for the characters to enhance the interaction the characters all have emotionally expressive animations. Garau et al. [28] point out that avatars that do not exhibit appropriate emotional expression during

emotionally charged conversation can be detrimental to an interaction. The problem with animated avatars is that they can be very complex to use if users have to directly control the avatars animation. Vilhjálmsson and Cassell [29]) have shown that users find controlling animated avatars difficult and their experience and interaction is improved if they use an avatar whose behaviour is controlled autonomously. We therefore have an autonomous model of affective animation for our avatars based on the affective states detected in users' text input. These detected affective states control the animation of the user avatars using Demeanour expressive animation framework [30].

Fig. 3. Four actors in the green room

Demeanour makes it possible for our characters to express the affective states detected by EMMA. When EMMA detects an affective state in a user's text input, this is passed to the demeanour system attached to this user's character and a suitable emotional animation is produced. The animation system is based around a set of short animation clips, each of which is labelled with one or more affective states. Each clip only affects an individual part of the body (torso, legs, arms) and thus several clips can be easily combined at the same time. When a new affective state is received a new set of clips is chosen at random from the clips labelled with that state and these new clips are combined together to produce a new animation. Every few seconds the set of clips used is varied at random to produce a new animation, but one which has the same affective state as before. This allows us to produce varied behaviour over long time periods. The animation system also implements affective decay. Any affective state will eventually revert to a neutral state if it is not replaced by a new one.

Another feature of the animation system is that characters can produce affective responses to the states of other characters. If a character produces a strong affective state then other characters will also produce a milder response. Each character has a profile which specifies how it responds to the behaviour of each other character. This makes it possible to implement different responses for different characters. For example, two characters with a positive relationship may empathise with each other, when one is unhappy so is the other. On the other hand if two characters have a negative relationship then one might gloat at the other's unhappiness, and therefore display happiness.

Fig. 4. Emotional believable agents on stage

Demeanour generates a number of output affective states, which are used to select the animation clips. Each output state is a weighted sum of a number of input factors. The primary input factor is the affective state as detected from the input text, this always has weight 1. The inputs also include the states of other characters, with lower weights. So for example, the output state "happiness" depends on the input "happiness", but also on the "happiness" and "sadness" values of other characters. The weights of the other characters states are contained in the character's profile. The profile consists of a separate set of weights for each other character in the scenario. This makes it possible to respond differently to each character. For example, if two characters, A and B have a poor relationship and A is angry with B, B might respond by being angry back. On the other hand if B's parents were angry then B might be sad or submissive. At any time each character has a single focus of social attention (which is itself another character), determined by the character's direction of gaze (which is itself determined by an animated gaze model). In order to generate animation the first step is to update the input states based on any text typed in. Next the states of the current focus of attention are fetched. These are multiplied by the weights of given by the profile specific to the focus of attention and added to the input emotion to produce the output state.

3 Affect Via Metaphor – Theoretical Analysis in E-Drama Transcripts

The preliminary analysis of transcripts from the user-testing that we have performed (see below) has shown or suggested a number of things concerning metaphor, as follows.

It has provided additional evidence for the importance of one of the principles we have put forward in our theoretical work on metaphor concerning the meaning of

metaphorical utterances. This concerns various types of information that we propose are transferred by default in metaphor generally, irrespective of the particular metaphorical view or views taken by the utterance (e.g., whether it views a mind as a physical container or a company as an animal). We have proposed in other work [24, 25] that – amongst various other types of information – emotional attitudes, value judgments and propositions concerning reduced functionality are transferred from source to target. We call the particular transfer principles involved view-neutral mapping adjuncts (VNMAs). Much if not most of the discourse contribution of a metaphorical utterance can be conveyed by the action of VNMAs, and this is borne out by many of the examples of metaphor in the e-drama transcripts.

Indeed, the conveying of emotional attitudes and the description of emotional states seems to be done more commonly in our genre via VNMAs than by specific metaphorical views of emotion such as those discussed by Kövecses [18] (an example of such a view is ANGER AS HEATED LIQUID).

These points are highly significant in terms of metaphor theory, because

(a) the central role of transfer of emotion and value judgments in metaphor has often been recognized but has not previously been properly systematized in theories,

(b) VNMAs factor out metaphorical-view-neutral transfers that other researchers have cast as being aspects of specific metaphorical views – we are therefore able to propose much simpler accounts of specific metaphorical views.

The types of metaphor that have arisen in e-drama transcripts sometimes resist pigeonholing neatly into the array of metaphorical views that have been suggested in the literature, so that new views are suggested. For example, we suggest a general view whereby INTERPERSONAL EFFECTS are viewed as PHYSICAL EFFECTS. Particular special cases of this have been suggested in the literature, and some cases of the new view fit the much more general CAUSES AS FORCES view [18], but our new intermediate view leads to a more focussed and appropriate analysis.

Many authors have struggled with the task of relating metaphor to, and distinguishing it from, other types of figurative language, such as metonymy and hyperbole. A number of e-drama examples lend weight to our suspicion that the distinction between metaphor and other figures is yet more troublesome than has been assumed so far. For example, many insults rely on casting someone as insane: this can potentially be analysed either as just hyperbole or as hyperbolic metaphor. Our work is leading to new light being thrown on such analytical difficulties, which have consequences for automated processing approaches as well as for theory.

Finally, the e-drama examples throw into relief the importance of further clarifying how size attributions, for instance via the adjectives “big” and “little”, contribute to metaphorical descriptions, as mentioned above. We are studying the extent to which the effects could be produced by VNMAs, or whether on the other hand they rest on specific metaphorical views.

4 User Testing

We conducted a two-day pilot user test with 39 secondary school students in May 2005, in order to try out and a refine a testing methodology. The aim of the testing was primarily to measure the extent to which having EMMA as opposed to a person

play a character affects users' level of enjoyment, sense of engagement, etc. We concealed the fact that EMMA was involved in some sessions in order to have a fair test of the difference that is made. We obtained surprisingly good results. Having a minor bit-part character called "Dave" played by EMMA as opposed to a person made no statistically significant difference to measures of user engagement and enjoyment, or indeed to user perceptions of the worth of the contributions made by the character "Dave". Users did comment in debriefing sessions on some utterances of Dave's, so it was not that there was a lack of effect simply because users did not notice Dave at all. Also, the frequencies of human "Dave" and EMMA "Dave" being responded to during the improvisation (sentences of Dave's causing a response divided by all sentences said by "Dave") are both roughly around 30%, again suggesting that users notice Dave. Additionally, the frequencies of other side-characters being responded to are roughly the same as the "Dave" character – "Matthew": around 30% and "Elise": around 35%.

Furthermore, it surprised us that few users appeared to realize that sometimes Dave was computer-controlled. We stress, however, that it is not an aim of our work to ensure that human actors do not realize this.

More extensive user testing at several Birmingham secondary schools has been conducted recently (up to September 2006). We have conducted an initial evaluation of the quality of EMMA's determinations about emotion during these school testing sessions, by comparing EMMA's determinations during one of the School Bullying improvisations with emotion labels later assigned offline by two members of our team (not centrally involved in the development of EMMA's algorithms). The humans were constrained to use the set of emotion labels that EMMA uses. It is still unclear how to conduct such an evaluation, because the conscious thoughts of a human annotator (labeller) about emotions revealed by an utterance in an offline labelling task may be different from the emotions the annotator would have unconsciously understood during an online e-drama session, and EMMA only surmises an emotion when there is strong evidence whereas a human labeller may proceed on a different level of evidence. Also, it transpires that there was mediocre agreement between the two human labellers, and the task is artificial for them because they might normally have assigned an emotion outside the prescribed set of 25. To compare the labelling of the two human labellers to each other and to the labelling by EMMA, we used the often-used *kappa* statistic of Carletta [31]. It is a measure of the pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement. The statistic, K, is calculated as $K= (P(A)-P(E))/ (1-P(E))$ where P(A) is the proportion of times two coders agree and P(E) is the proportion of times we would expect them to agree if they categorized randomly. A value of at least 0.6 – 0.8 is generally required by researchers looking for good inter-annotator agreement. We calculated K for each pair among the three labellers (EMMA and two humans). The inter-human K was only 0.32, and so it is not surprising that the EMMA/human values were only 0.32 again and 0.23. However, we also performed a modified comparison in which the emotion labels were conflated to three (positive, negative, neutral) by grossly lumping together, for example, the labels deemed positive by our team. We then got a human/human K of 0.65, and EMMA/human values of 0.55 and 0.42. The latter are not good values, but they at least give grounds for hope that with further refinement of our affect detection we can come near the rather low human/human level of agreement.

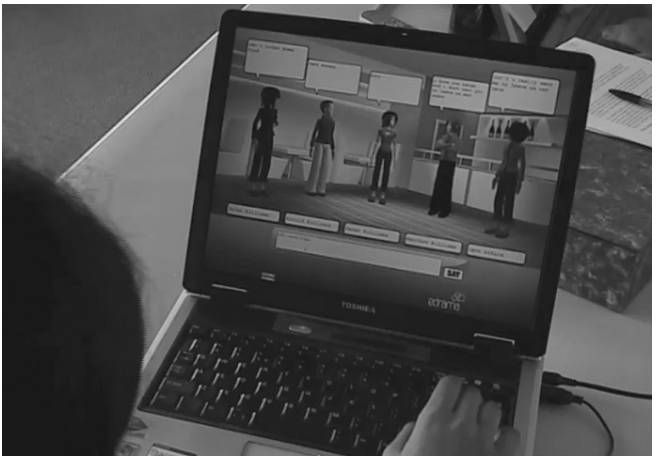


Fig. 5. E-drama user testing with school children

Inspection of the transcripts collected indicates that EMMA usefully pushed the improvisation forward on various occasions. Fig. 6. shows an example about how EMMA contributed to the drama improvisation in Crohn's disease scenario. Briefly, in Crohn's disease scenario, Peter has had a terrible bowel disease – Crohn's disease – since he was 15. Crohn's disease attacks the wall of the intestines and makes it very difficult to digest food properly. Peter has the option to undergo surgery (ileostomy) which will have a major impact on his life, but he will carry a bag outside his stomach for the rest of his life. So he needs his family (Dad: Arnold, Mum: Janet, brother: Matthew) and his close friend (Dave) to help him make a decision about having another operation or not. He decides to take them out to a restaurant to talk about it. His dad (Arnold) loves him but always feels uncomfortable and embarrassed when Peter talks about his disease in public. Peter's mum cares about him very much, sometimes worries too much about Peter. Matthew is totally against the operation idea and believes Peter will get bullied because of the side-effect of the operation. Dave, Peter's best mate, loves Peter and his family and tries to mediate the discussion. In the example transcript shown in Fig. 6, Dave was played by EMMA, which successfully led the improvisation on the desirable track. In another scenario (school bully scenario) that we have used for the user testing, example transcripts are also shown that the EMMA character has helped to push the improvisation forward.

The experimental methodology used in the testing is as follows, in outline. Subjects are 14-16 year old students at local Birmingham schools. Forty students are chosen by each school for the testing. Four two-hour sessions take place at the school, each session involving a different set of ten students. In a session, the main phases are as follows: an introduction to the software; a First Improvisation Phase, where five students are involved in a School Bullying improvisation and the remaining five in a Crohn's Disease improvisation; a Second Improvisation Phase in which this assignment is reversed; filling out of a questionnaire by the students; and finally a group discussion acting as a debrief phase. For each improvisation, characters are pre-assigned to specific students. Each Improvisation Phase involves some preliminaries followed by ten minutes of improvisation proper.

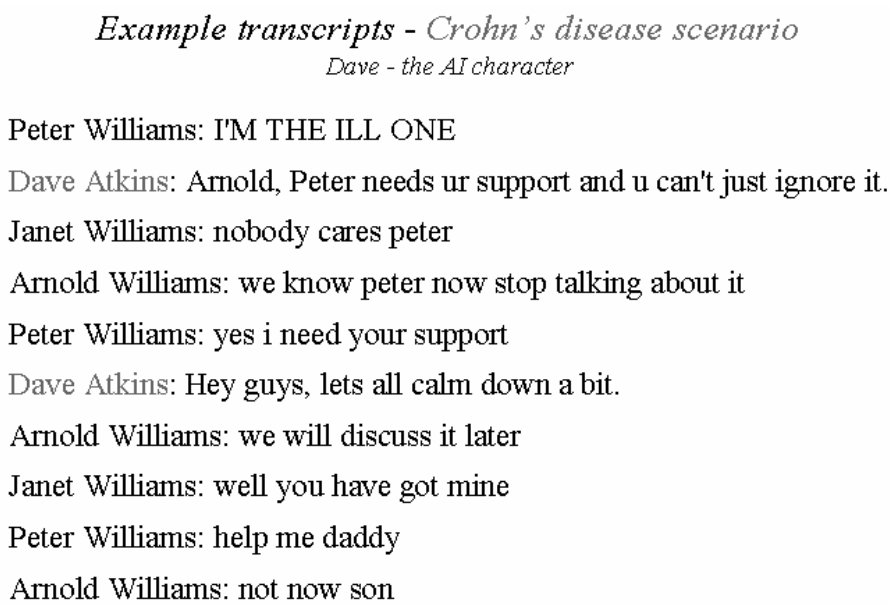


Fig. 6. Part of an example transcript produced in the user testing in Swanshurst School, Birmingham. 28th June 2006.

In half of the SB improvisations and half of the CD improvisations, the minor character Dave is played by one of the students, and by EMMA in the remaining. When EMMA plays Dave, the student who would otherwise have played him is instructed to sit at another student's terminal and thereby to be an audience member. Students are told that we are interested in the experiences of audience members as well as of actors. Almost without exception students have appeared not to have suspected that having an audience member results from not having Dave played by another student. At the end of one exceptional session some students asked whether one of the directors was playing Dave.

Of the two improvisations a given student is involved in, exactly one involves EMMA playing Dave. This will be the first session or the second. This EMMA-involvement order and the order in which the student encounters SB and CD are independently counterbalanced across students.

The questionnaire is largely composed of questions that are explicitly about students' feelings about the experience (notably enjoyment, nervousness, and opinions about the worth of the dramatic contributions of the various characters), with essentially the same set of questions being asked separately about the SB and the CD improvisations. The other data collected are: for each debrief phase, written minutes and an audio and video record; notes taken by two observers present during each Improvisation Phase; and automatically stored transcripts of the sessions themselves, allowing analysis of linguistic forms used and types of interactivity. To date only the non-narrative questionnaire answers have been subjected to statistical analysis, with the sole independent variable being the involvement or otherwise of EMMA in improvisations.

5 Conclusion and Ongoing Work

We have implemented a limited degree of affect-detection in an automated bit-part character in an e-drama application, and fielded the character successfully in pilot user-testing. Although there is a considerable distance to go in terms of the practical affect-detection that we plan eventually to implement, the already implemented detection is able to cause contributions by the automated character that are reasonably appropriate to the discourse.

Additionally, E-drama provides a platform for participants to engage in focused discussion around emotionally charged issues. This new prototype provides an opportunity for the developers to explore how emotional issues embedded in the scenarios, characters and dialogue can be represented visually without detracting from the learning situation.

Our project makes a contribution to the issue of what types of automation should be included in interactive narrative environments, and as part of that the issue of what types of affect should be detected (by directors, etc.) and how. Moreover, our project also makes a contribution to the development of automatic understanding of human language and emotion in human computing. The generation of emotional believable animations based on the detected affective states contributes to the ease and innovative user interface in e-drama, which leads to high-level user engagement and enjoyment.

Our remaining work on affect detection within the project will be on the metaphorical aspects of it, both in the sense of developing our theoretical ideas about metaphor further in the light of the e-drama data and in the sense of implementing certain limited forms of metaphor processing that can reveal useful extra hints about the affect that is present.

Future work could also include the exploration of automated bit-part characters to fully develop a non-human director. Additionally tools to enable participants to replay the role-plays have been considered. These could enable further reflection and group discussion, allowing for comparisons of sessions between different groups of learners. Replays could even be altered to adjust the emotional states of each character and generate different online ‘performances’, which could create emotionally rich experiences for audiences as well as participants.

Acknowledgments. This work is supported by grant RES-328-25-0009 from the ESRC under the ESRC/EPSRC/DTI “PACCIT” programme. We are grateful to Hi8us Midlands Ltd, Maverick Television Ltd, BT, and our colleagues Z. Wen, R. Agerri, W.H. Edmondson and S.R. Glasbey. The work is also partially supported by EPSRC grant EP/C538943/1.

References

1. Zhang, L., Barnden, J.A., Hendley, R.J. & Wallington, A.M. 2006a. Exploitation in Affect Detection in Open-ended Improvisational Text. In *Proceedings of Workshop on Sentiment and Subjectivity* at COLING-ACL 2006, Sydney, July 2006.
2. Zhang, L., Barnden, J.A., Hendley, R.J. & Wallington, A.M. 2006b. Developments in Affect Detection in E-drama. In Proceedings of EACL 2006, *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, Trento, Italy. pp. 203-206.

3. Picard, R.W. 2000. *Affective Computing*. The MIT Press. Cambridge MA.
4. Ortony, A., Clore, G.L. & Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge U. Press.
5. Prendinger, H. and Ishizuka, M. Simulating Affective Communication with Animated Agents. In *Proceedings of Eighth IFIP TC.13 Conference on Human-Computer Interaction*, Tokyo, Japan, pp.182-189.
6. Wiltschko, W. R. Emotion Dialogue Simulator. eDrama learning, Inc. eDrama Front Desk.
7. Mehdi, E. J., Nico P., Julie D. and Bernard P. 2004. Modeling Character Emotion in an Interactive Virtual Environment. *Proceedings of AISB 2004 Symposium: Motion, Emotion and Cognition*. Leeds, UK.
8. McCrae, R.R. and John, O.P. 1992. An Introduction to the Five Factor Model and Its Application. *Journal of Personality*, 60, 175-215.
9. Gratch, J. and Marsella, S. A Domain-Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*. Vol 5, Issue 4, pp.269-306.
10. Egges, A., Kshirsagar, S. & Magnenat-Thalmann, N. 2003. A Model for Personality and Emotion Simulation, In *Proceedings of Knowledge-Based Intelligent Information & Engineering Systems (KES2003)*, Lecture Notes in AI. Springer-Verlag: Berlin.
11. Elliott, C., Rickel, J. & Lester, J. 1997. Integrating Affective Computing into Animated Tutoring Agents. In *proceedings of IJCAI'97 Workshop on Intelligent Interface Agents*.
12. Aylett, R.S., Dias, J. and Paiva, A. (2006) An affectively-driven planner for synthetic characters. In *Proceedings of ICAPS 2006*.
13. Mateas, M. 2002. Ph.D. Thesis. Interactive Drama, Art and Artificial Intelligence. School of Computer Science, Carnegie Mellon University.
14. Zhe, X. & Boucouvalas, A. C. 2002. Text-to-Emotion Engine for Real Time Internet Communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs*, Staffordshire University, UK, pp 164-168.
15. Boucouvalas, A. C. 2002. Real Time Text-to-Emotion Engine for Expressive Internet Communications. In *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*. G. Riva, F. Davide and W. IJsselsteijn (eds.) 305-318.
16. Craggs, R. & Wood. M. 2004. A Two Dimensional Annotation Scheme for Emotion in Dialogue. In *Proceedings of AAAI Spring Symposium: Exploring Attitude and Affect in Text*.
17. Fussell, S. & Moss, M. 1998. Figurative Language in Descriptions of Emotional States. In S. R. Fussell and R. J. Kreuz (Eds.), *Social and cognitive approaches to interpersonal communication*. Lawrence Erlbaum.
18. Kövecses, Z. 1998. Are There Any Emotion-Specific Metaphors? In *Speaking of Emotions: Conceptualization and Expression*. Athanasiadou, A. and Tabakowska, E. (eds.), Berlin and New York: Mouton de Gruyter, 127-151.
19. Watson, D. & Tellegen, A. 1985. Toward a Consensual Structure of Mood. *Psychological Bulletin*, 98, 219-235.
20. Ekman, P. 1992. An Argument for Basic Emotions. In *Cognition and Emotion*, 6, 169-200.
21. Werry, C. 1996. Linguistic and Interactional Features of Internet Relay Chat. In *Computer-Mediated Communication: Linguistic: Social and Cross-Cultural Perspectives*. Pragmatics and Beyond New Series 39. Amsterdam: John Benjamins, 47-64.
22. Briscoe, E. and J. Carroll. 2002. Robust Accurate Statistical Annotation of General Text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Gran Canaria. 1499-1504.
23. Heise, D. R. 1965. Semantic Differential Profiles for 1,000 Most Frequent English Words. *Psychological Monographs*. 70 8:(Whole 601).

24. Barnden, J., Glasbey, S., Lee, M. & Wallington, A. 2004. Varieties and Directions of Inter-Domain Influence in Metaphor. *Metaphor and Symbol*, 19(1), 1-30.
25. Barnden, J.A. Forthcoming. Metaphor, Semantic Preferences and Context-sensitivity. Invited chapter for a Festschrifft volume. Kluwer.
26. Esuli1, A. and Sebastiani, F. 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining. In Proceedings of EACL-06, *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, IT. pp. 193-200.
27. Sharoff, S. 2005. How to Handle Lexical Semantics in SFL: a Corpus Study of Purposes for Using Size Adjectives. *Systemic Linguistics and Corpus*. London: Continuum.
28. Garau, M., Slater,M., Bee, S. and Sasse, M.A. 2001. The impact of eye gaze on communication using humanoid avatars. *Proceedings of the SIG-CHI conference on Human factors in computing systems*, March 31 - April 5, 2001, Seattle, WA USA, 309-316.
29. Vilhjálmsson, H. & Cassell, J. 1998. BodyChat: Autonomous Communicative Behaviors in Avatars. *Proceedings of ACM Second International Conference on Autonomous Agents*, May 9-13, Minneapolis, Minnesota.
30. Gillies, M., Crabtree, I.B. and Ballin, D. 2006. Individuality and Contextual Variation of Character Behaviour for Interactive Narrative. In *Proceedings of the AISB Workshop on Narrative AI and Games*.
31. Carletta, J. 1996. Assessing Agreement on Classification Tasks: The Kappa statistic. *Computational Linguistics*, 22 (2), pp.249-254.
32. Pantic, M. Pentland, A. Nijholt, A. and Huang, T. 2006. Human Computing and Machine Understanding of Human Behavior: A Survey. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 239-248.
33. Cohn, J.F. 2007. Foundations of human-centered computing: Facial expression and emotion. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)* Hyderabad, India.
34. Nogueiras et al. 2001. Speech emotion recognition using hidden Markov models. In *Proceedings of Eurospeech 2001*, Denmark.
35. Weizenbaum, J. 1966. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9(1): 36-45.

Author Index