J Multimodal User Interfaces (2012) 6:117–141
DOI 10.1007/s12193-012-0093-9

ORIGINAL PAPER

A generic framework for the inference of user states in human computer interaction

How patterns of low level behavioral cues support complex user states in HCI

Stefan Scherer · Michael Glodek · Georg Layher · Martin Schels · Miriam Schmidt · Tobias Brosch · Stephan Tschechne · Friedhelm Schwenker · Heiko Neumann · Günther Palm

Received: 20 June 2011 / Accepted: 6 March 2012 / Published online: 5 April 2012

Abstract The analysis of affective or communicational states in human-human and human-computer interaction (HCI) using automatic machine analysis and learning approaches often suffers from the simplicity of the approaches or that very ambitious steps are often tried to be taken at once. In this paper, we propose a generic framework that overcomes many difficulties associated with real world user behavior analysis (i.e. uncertainty about the ground truth of the current state, subject independence, dynamic realtime analysis of multimodal information, and the processing of incomplete or erroneous inputs, e.g. after sensor failure or lack of input). We motivate the approach, that is based on the analysis and spotting of behavioral cues that are regarded as basic building blocks forming user state specific behavior, with the help of related work and the analysis of a large HCI corpus. For this corpus paralinguistic and nonverbal behavior could be significantly associated with user states. Some of our previous work on the detection and classification of behavioral cues is presented and a layered architecture based on hidden Markov models is introduced. We believe that this step by step approach towards the understanding of human behavior underlined by encouraging preliminary results outlines a principled approach towards the development and evaluation of computational mechanisms for the analysis of multimodal social signals.

S. Scherer (✉)
Institute for Creative Technologies, University of Southern California, Los Angeles, USA
e-mail: scherer@ict.usc.edu

M. Glodek · G. Layher · M. Schels · M. Schmidt · T. Brosch · S. Tschechne · F. Schwenker · H. Neumann · G. Palm
Institute of Neural Information Processing, Ulm University, Ulm, Germany

Keywords Multimodal information fusion · Nonverbal behavioral cue detection · Layered architecture · Fuzzy data processing · Human-computer interaction

1 Introduction

Currently human-computer interaction (HCI) typically takes place on a rather crude explicit question-answer level, whereas human-human interaction is multifaceted, consisting of manifold interactive feedback loops between interlocutors, comprising social components (e.g. display rules, social state), moods, feelings, personal goals, nonverbal and paralinguistic conversation channels and more [14, 49, 55]. In order to close this gap, it is crucial for a machine to perceive and understand the user's current interaction and affective state, as well as it is necessary to register the user's social signals, which are composed of multiple behavioral cues [123]. Most of the research aiming at recognizing the user's state focuses on the recognition of emotions [21], often the so called "big six" introduced by [23] and [31]. As stated in [89], traditional theory on emotion is often oriented on extreme or full-blown emotions that rarely occur in HCI.

Comparatively, little effort is spent to investigate user states that rather concern the user's attitude towards the interaction or define the current interaction's state; these states are shared between the interlocutors (e.g. involvement, agreement, understanding, etc.) be they human or artificial. These states influence the conversation immensely and are not reduced to the user's internal emotional state. In order to be able to assess such states from realistic interactions without known ground truths, it is necessary to conduct human perception tests. In such tests uninformed subjects are invited to report what they perceive, while watching or listening to these interactions with the help of compiled lists, or other annotation tools. Humans infer the interaction state

from the context information, e.g. the history of the dialog and the scenery, and the observable social signals composed by paralinguistic events and nonverbal communication cues. Several current research projects aim towards understanding those behavioral cues and social signals, e.g. FASTNET[1] (Focus on Action in Social Talk: Network-Enabling Technology), SSPNET[2] (Social Signal Processing Network) or the SFB/TRR 62 Companion Technology for Cognitive Technical Systems project.[3]

Motivated by this, the present paper introduces a proposal to tackle the problem. The presented work should be regarded as a pilot study providing a first framework for future developments incorporating state of the art approaches able to handle the discussed requirements from a technical multimodal signal processing and fusion view.

We start with the analysis of a set of labels, describing user dispositions in HCI based on a realistic HCI corpus (as introduced in [99]). The labels comprise categories of different complexity: several are directly inferable (e.g. a subject is laughing), while others are only accessible, when provided with context of the interaction—even for human annotators. In this context, the aim is to find statistical correlations leading from observations and behavioral cues to subject states.

We then report some of our relevant previous work with respect to necessities of application-driven social signal processing. On the one side it is important to be able to spot behavioral cues in interactions that appear on a short-term scale and on the other side the integration of their detections over short as well as longer temporal periods to be able to infer social signals and communicational states is crucial.

In order to support such a system functionality realtime processing capabilities are one requirement. For example, laughter (which forms a basic building block in our approach) in continuous conversational input streams needs to be spotted instantaneously to steer the further conversation. Further, as social signals are often conveyed in multiple modalities, the integration of multimodal inputs at different timescales impose hard constraints on the processing capabilities. In order to be able to integrate these findings it is essential to be able to handle uncertain information (e.g. quantifying the risk of lower level classifications), to compensate possible sensor failure/outage (e.g. facial features are not visible all the time), and to detect dynamics and patterns within the occurrences of nonverbal communication elements.

The remainder of the paper is organized as follows: Sect. 2 introduces the requirements for a framework architecture from findings in the literature and related work, including: possible target states, nonverbal and paralinguistic basic building blocks, suitable machine learning architectures to spot these cues, as well as approaches to integrate them. The proposed perspective to look at HCI is supported by a layered annotation of the PIT[4] corpus [111], a multiparty HCI dataset, which is presented in detail in Sect. 3. Furthermore, within this section statistical relations between lower level observations (e.g. behavioral cues, speech acts), and complex user states are examined. Requirements and concrete examples of machine learning approaches from our previous work are presented in Sect. 4 and show the capabilities of current state of the art machine learning to spot and integrate behavioral cues. Studies evaluating these approaches in the context of HCI are presented to provide a deeper insight. In Sect. 5 the proposed framework is discussed and Sect. 6 concludes the paper.

2 Framework requirements and related work

2.1 Affective computing targets and resources

The theory and concepts on emotion, as it is stated in [88], are "parts of a folk theory inherited from human shepherding and farming ancestors", and the understanding of it became some sort of "common sense" to computer scientists and engineers working on affective computing in HCI. Some researchers, however, have called it a "very confused and confusing field of study" [76], which indicates the difficulty that has to be dealt with by approaching this field of science. Furthermore, in [88] it is mentioned that the prominent question: "What is an emotion?" remains still under debate and a consensus is far from being found.

Early affective computing development has mainly focused on typical holistic psychological emotion theories [23, 48]. However, these classic approaches do not easily fit onto realistic HCI scenarios, as they always comprise an enormous spectrum of possible emotions,[5] that are not likely to happen in daily human life let alone in realistic HCI[6] [89].

Early affective computing corpora often included only a limited set of acted emotional—often unimodal—clips comprising typically up to six or seven emotions as target states (e.g. anger, happiness, sadness, boredom, surprise) [4, 11, 128]. The ecological validity of such corpora

[1] http://www.tcd.ie/slscs/research/projects/current/fastnet.php

[2] http://sspnet.eu/

[3] http://www.sfb-trr-62.de/

[4] Competence center for Perception and Interactive Technologies (PIT) at Ulm University.

[5] As mentioned in [89]: "As psychologists use the term, it includes the euphoria of winning an Olympic gold medal, a brief startle at an unexpected noise, unrelenting profound grief, the fleeting pleasant sensations from a warm breeze, cardiovascular changes in response to viewing a film, the stalking and murder of an innocent victim, lifelong love of an offspring, feeling chipper for no known reason, and interest in a news bulletin."

[6] Even though, some may argue that hot anger is a common feeling one might have towards the ineptitude of the operating system.

of course suffers from the fact that acted representations are biased and not entirely natural. No context of the situation is given and the expression is entirely a product of the imagination of the actors, who are often laymen. Hence, more intense and prototypical portrayals are expected to be found in those corpora; in contrast to observations observed in real HCI or human-human interaction. It has to be expected that more obvious cues, such as the loudness of the speech are exaggerated, whereas more subtle cues, such as the voice quality, are underrepresented.[7] However, the main advantage often outweighs the disadvantages; acted data it is comparably cheap to gather and the target labels are known. That said, it is clear that the acted approach, relying on short clips of stereotypical portrays, lacks the presence of non-verbal and paralinguistic communication elements, such as laughter, moans, subtle gestures or conversational dynamics.

Realistic and spontaneous recordings on the other side, have the highest validity and are more representative than acted and induced data. However, they often suffer from some drawbacks, such as the limited number of speakers and recordings, or noisy environments [92]. Furthermore, it is difficult to assess the participants' true affective state, which is why it is important to utilize experienced human annotators interpreting the situation. The lack of the known underlying ground truth, adds uncertainty, subjective interpretations and opinions to the task.

Often the source of naturalistic or realistic expressive recordings are TV shows or live news broadcasts, as for the corpus described in [27]. These recordings, however, might put the ecological validity into question due to self-presentation effects or display rules that apply strongly in situations of public behavior [92]. In [14] it is further mentioned that any form of intrusion using recording devices or sensors into the normal life of the subject already does change the behavior. This phenomenon is closely related to Labov's observer's paradox which could also be called the "corpus maker's paradox" [14] in this context. Hence, it could be argued that only recordings in which the subjects got entirely used to the fact that they are being recorded are valid. However, such a corpus is very expensive to record and requires a lot of time to acquire and analyze (i.e. recording and annotation).

A wizard of Oz system capable of simulating typical situations of a target application area is often a good trade-off. The ecological validity in such recordings is of course limited to the simulated system and the interaction with it, but renders a compromise between realism and a good cost/performance ratio. In Sect. 3, we base our analysis on a corpus that was recorded in a wizard of Oz based scenario, in which the participants were believed to interact with a fully automatic machine.

In most of the affective computing studies emotional or affective states are considered to be personal and most of the annotations only allow the labeling of personal and internal emotional states, using basic sets such as "big six" emotions (mixed annotations are allowed sometimes), or affective dimension annotations [27]. The analysis of those is complicated by several social factors such as display rules, social standing, personality factors (personality traits and attribution) and the like [20, 22].

Not only does the automatic analysis of internal emotional states invade the user's privacy, but it also misses a very important factor of communication, namely the shared communicational states, that are essential for the understanding and the participation in interactions. A lot of research nowadays goes into that direction and tries to understand the factors and phenomena that are observable within the interlocutors' shared space. In the following, we mention a few of the works analyzing such communicational states.

In [15] for example, moments of high engagement[8] of conversations are analyzed using movement information of four interlocutors interacting in a free and uncontrolled manner. Others call these moments of excitement, interest and heated discussions in conversations "hotspots" [130]. In general, participant involvement is an important feature for the assessment of the quality of conversations and is often characterized by various paralinguistic features of speech and various other behavioral cues [72, 73, 80].

States of accommodation,[9] which is established over a perception-behavior link [34, 60], indicates reduced social distance, rapport and affiliation [60, 110]. It establishes comprehension and understanding [34], as well as signals agreement or disagreement [117, 122].

The above mentioned communicational states and dispositions are also relevant in the scenario of HCI and also in social interaction, as they represent a shared quality of interaction; we annotated a set of those within the PIT corpus, e.g. interest, disagreement, or agreement, as reported in Sect. 3.1 and in [95, 99]. However, we are well aware that the conducted annotations of the PIT corpus are no where near exhaustion and could be extended in the future.

2.2 Behavioral cues as basic building blocks

In order to assess the underlying users' affective state or communicative disposition in typical HCI scenarios, human annotators can only indirectly infer them using background

[7] One way of dealing with these stereotypical expressions, could be the use of perception tests to filter exaggerations and unnatural behavior out [11, 92].

[8] Referred to by the Japanese term "moriagari" [15].

[9] Other often synonymously used terms in the literature are: Mimicry [80], entrainment [10], alignment [82], or synchrony [29].

information (i.e. the context of the interaction) and observations of the participants' behaviors (e.g. behavioral cues and their combination to form a social signal [123]). Wherein, gestures, mimics and the tone of voice contribute almost equally to the inference of affective states as mentioned in [78].

The constant identification and interpretation[10] of nonverbal communication cues or social signals, which is often done by humans unconsciously, to assess opinions on the qualities of an interaction, has been and is a vibrant field of research [123]. Behavioral cues comprise but are not limited to mimics (e.g. smiles, lifting of eyebrows etc.), gestures (e.g. hand gesturing etc.), posture of the body, moans, grunts, back-channelling behavior, laughter, pauses, and other prosodic cues [8, 12, 13, 32, 48, 55, 126]. In the following, we mention some relevant examples of combinations and occurrences of such cues found in the literature that are closely related to various communication states, such as agreement, engagement, or interest.

Agreement and disagreement, as analyzed in [8], is conveyed by various signals; agreement on the one side is portrayed by, e.g. laughter, head nod, and mimicry; disagreement on the other side by, e.g. head shake, scratching the head, and gaze aversion. Additionally, in [122] prosodic synchrony could be identified as a cue for spotting agreement.

In [84] the importance of laughter as a communicative discourse element, and as a prominent social signal almost exclusive to social encounters is underlined and the acoustic characteristics are thoroughly analyzed. In particular, laughter is an indication for the positive perception of a discourse element, or an indication for uncertainty considering nervous or social laughter [12]. Overall, laughter is necessary for "healthy" communication and occurrences can be used to measure engagement or interest in interaction [50, 53, 61, 120]. Laughter is acoustically highly variable and is expressed in many forms; giggles, exhaled or inhaled laughs, or even snort like laughter exist [3, 12]. In Sect. 4.3.1, a short summary on a thorough analysis to spot laughter in running speech is given.

Involvement or active interest, are often expressed by direct gaze, frequent gesturing, smiles, or other nonverbal expressions [2]. In [72], average movement, speaking rate, and fundamental frequency (f_0) statistics (i.e. median and range) could be identified as indicators for participants' involvement.

Other paralinguistic cues, such as voice quality (e.g. breathy, tense, and modal voices) and their dynamic use in spoken language can reveal useful information on a speaker's attitude, mood and affective state. As voice quality is thought to be very important for the assessment of the affective state of the speaker [65, 131, 132]. Each speaker possesses his own habitual voice quality which may or may not relate closely to Laver's description of modal voice (i.e. efficient vocal fold vibration with little or no aspiration noise) [63]. For instance, a given speaker's habitual voice quality may be inherently breathy, whereas for another speaker the same quality of breathiness may be quite a deviation from their habitual voice quality. However, the deviations from the speaker's baseline bear most information about his attitude or affective state, not the speaker's absolute voice quality setting.

Apart from more or less directly observable behavior, there is an additional set of non-observable physiological measures (i.e. electroencephalography, skin conductivity, electromyography, blood volume pulse, and respiration), that bear information about the affective state of a person [51]. Whereas, the collection of these measures somewhat intensifies the issues associated with the corpus maker's paradox [14], they still can improve the performance and advance the understanding of the underlying processes [90, 125].

It is important to bear in mind, that none of the above mentioned cues are sufficient as a single event (e.g. a single laughter may be uttered after a funny joke or because it is socially expected), to unambiguously convey communicational dispositions, and affective states. However, combinations of those multimodal cues following certain temporal patterns (that could of course vary) often suffice as mentioned in [8, 123]. This in turn is supporting a layered building block like structure which allows the spatiotemporal integration of behavioral cues, which forms the basis of the proposed overall architecture in this paper (see Fig. 1). We believe, that extracting patterns of these cues in social encounters or HCI represents a valuable strategy in order to tackle the difficult task of understanding and interpreting social signals and affective states. In the following requirements for automatic classifiers to spot behavioral cues, as well as their integration and extraction of patterns of them are discussed.

2.3 Challenges of social signal integration

In order to create a classifier to identify communicational user states and social signals, we follow the proposed spatiotemporal combination of behavioral cues to form social signals introduced in [123]. For this task an architecture as shown in Fig. 1, that processes the multimodal information on several levels simultaneously (e.g. audio and speech analysis, gesture analysis, posture analysis, and psychophysiology), is required. Classifiers and decisions, as well as subgoals, are combined using information fusion [57, 75, 118, 129].

One crucial aspect of the architecture is the integration or fusion of the various information channels. In principal, multiple fusion techniques may be applied, e.g. early feature level fusion, or decision level fusion [57]. Feature level

[10] Which of course may be error prone.

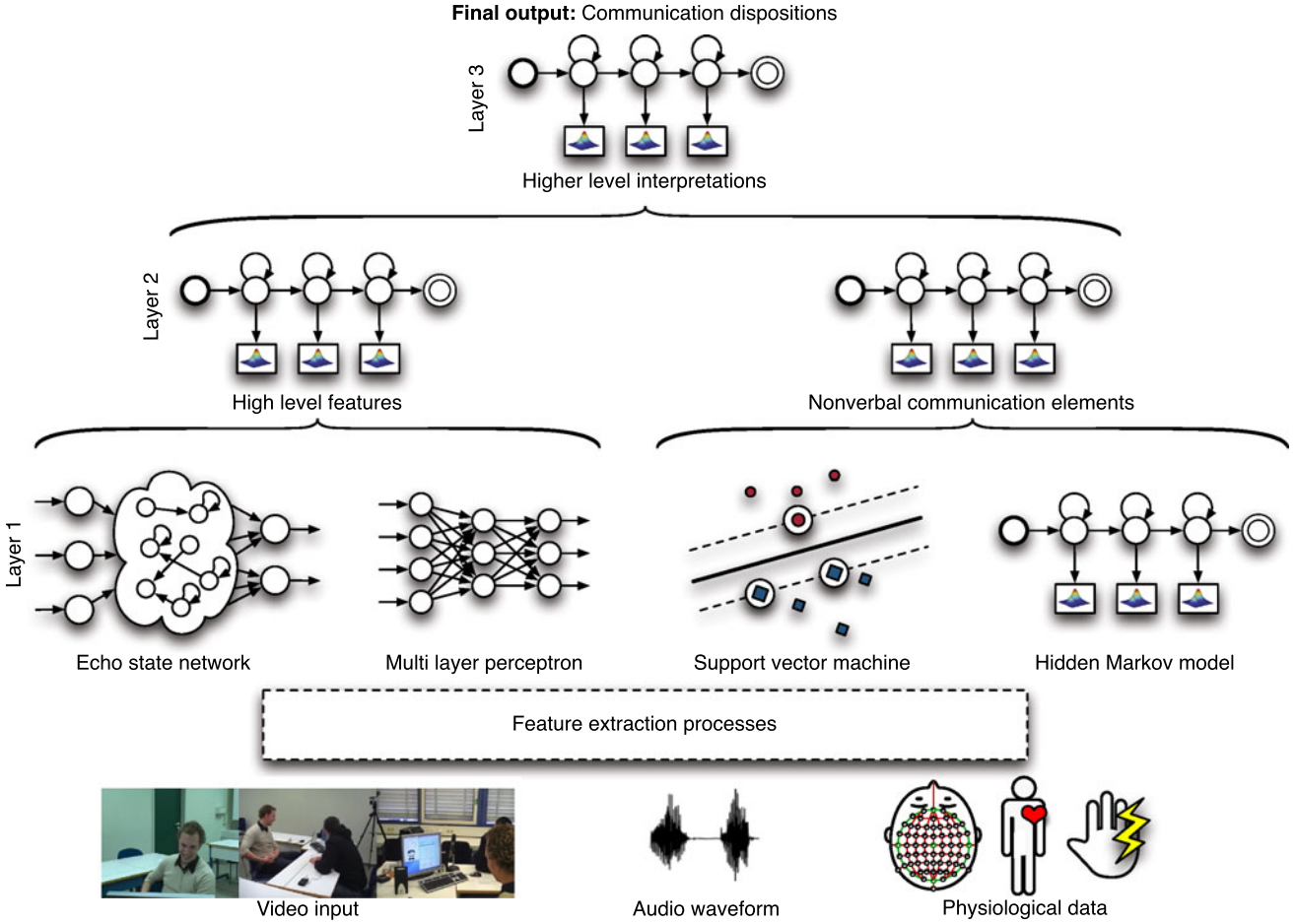

Fig. 1 Layout of a possible layered architecture integrating information from multimodal sensors (i.e. video, audio, and physiological sensory input) via classifiers detecting behavioral cues or various characteristics. The decisions are then propagated, with associated uncertainty, through the system to the sequence classifiers utilizing the temporal and dynamic patterns to infer complex user states in coarser time granularity

fusion approaches for example often lead to more accurate classification results; however, they are more difficult to handle due to varying sample rates or dynamics in the input data [18, 78, 93].

Combining classifiers is often a promising approach to improve the overall classifier performance [90, 91, 106]. Such a team of classifiers should be accurate and diverse [59]. While the requirement to the classifiers to be as accurate as possible is obvious, diversity roughly means that classifiers should not agree on misclassified data, and so multiple views to audio and visual data should be utilized to get an ensemble of diverse, complementary classifiers that can be aggregated to an overall efficient classifier system [25, 57].

Further, when combining multiple streams of data, some modalities may not be available at all times due to possible sensory failure in one modality or occlusions and obfuscations (e.g. hand in front of face; noisy environment).

Every classifier in the architecture should output probabilistic distributions over the intended categories of communicational states, or behavioral cues accompanied by certain confidence levels of the provided information. Classifiers above the bottom layer (i.e. the layer directly connected to the sensory input) are required to process the general fuzziness and uncertainty of data in this domain. Utilizing class membership values—rather than crisp class decisions—as input for the integration enables probabilistic models to improve robustness in the classification process, as well as allows the compensation of errors of the lower levels.

As argued in Sect. 2.2 and in [123], the dynamic appearances of nonverbal cues are assumed to characterize the different communicational states. Hence, a model for these states must be able to incorporate the temporal order of the detected cues.

As information is propagated from the lower layers to the higher layers the analyzed time scale of the observations becomes inherently more and more coarse. In order to pro-

cess time series, windowing techniques are required where at every step one decision is generated. Higher layers are fed with results and outputs of the former layer, which is causing the above mentioned effect. Hence, the amount of data available for learning the higher layers is inevitably smaller. On the other side, using such a model reduces the degrees of freedom compared to one bigger model that does not decompose into building blocks. This circumstance results in a more stable learning behavior and less data may be required to achieve a reliable classification [74, 75].

As no explicit theory, at least for the target domain HCI, and no rulesets describing the behavioral cues to form a social signal exist yet, one has to rely on data to build a model learning from examples. After training, such a model on the other side could be used to inform about typical spatiotemporal occurrences of behavioral cues that indicate social signals or user states.

In Sect. 4, an implementation of such an architecture, the layered HMM algorithm, will be explained in greater detail. The layered HMM approach covers a good part of the requirements for the inference of user or communication states from behavioral cues.

To summarize the challenges are listed in the following:

1. *Spatiotemporal information integration* over multiple modalities.
2. Integration of *uncertain information* on all layers (classifiers are error prone; ground truth is unknown; modalities may not be available at all times).
3. Architecture needs to be able to deal with *varying time scales and dynamics* as well as *occlusions and obfuscations*.

2.4 Requirements for behavioral cue spotting

According to our hypothesis low level observations, i.e. behavioral cues, are communicational signals made visible by the interlocutors, e.g. laughing or raising an eyebrow. Combinations of these cues often reveal insights into the mindset and affective state of the conversational partner or about the current communication state. As stated above (see Sect. 2.2) manifold of these signals exist and their temporal characteristics and patterns often correspond to relevant target states or social signals [123].

As the detected observations need to be lifted to a higher level context (e.g. raising eyebrow followed by laughing indicates surprise), additional requirements arise: crisp detections are, due to the large set of targets and possible confusions among them, often not sufficient within this scenario. More adequate would be the use of automatic approaches yielding fuzzy memberships or probability distributions over possible targets, such that the next layer which is integrating the information receives as much information as possible [103]. Using this approach it might be possible to correct for errors made in early stages of the classification in a later stage.

Algorithms which are able to render a distribution over classes such as feed-forward neural networks with a sigmoid output neuron [41], support vector machines (SVM) with probabilistic output [83] or probabilistic models incorporating distributions over classes [54, 116] suffice this requirement.

Due to individual differences and personality factors (compare Sect. 2.2), it is crucial that observations are correctly assigned to individuals, i.e. interlocutors participating in the conversation need to be identified and tracked correctly.

From the modeling view point it is crucial to design and use suitable features as well as classifiers: In order to fully access a modality, features need to be extracted from the corresponding data stream containing as much information about the target as possible; within the affective computing literature it turned out, that many features are suitable [6, 104]. Exploiting the temporal characteristics requires either features capturing the dynamics of the target [42–44] or algorithms capable of modeling the temporal structure of the features. Hidden Markov models (HMM) [85] or echo state networks (ESN) [45] are fitting candidates. HMM, however, do not provide a distribution over classes per se, but there are extensions to the model realizing this requirement [47].

The requirements are summarized in the following:

1. Distributions over targets or *fuzzy outputs* are required in order to be able to compensate for early detection errors.
2. Behavioral cues must be associated to the sender (*individual tracking* is required).
3. The design of *purpose tailored features* representing the target's dynamics and characteristics is crucial for efficient classification and understanding.
4. The design and use of *classifiers capable of modeling the dynamic structure* of target characteristics is of the essence.

3 The PIT corpus of German multiparty dialogs

The example data collection used for the evaluation of the hierarchical annotation scheme is the "PIT corpus of German multi-party dialogs" [111, 112], which is recorded using a wizard of Oz approach. The scenario is embedded in a restaurant search, and is composed of three dialog participants: two human subjects (U1 and U2), discussing their choice of a restaurant, and one computer (S) represented with an animated interface agent assisting them, in the best way possible. This human-human-machine interaction scenario bears rich paralinguistic contents and social signals

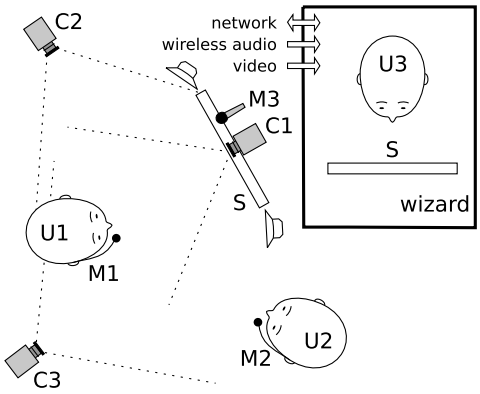

Fig. 2 Schematic view of the Wizard of Oz recording setup. The primary user marked as U1 interacts with the secondary user U2 and the System S. The wizard is located in another room and receives real time input from camera C4, and microphones M1 and M2. The subjects receive audiovisual output from the speakers and from the screen of S. Camera C1 records the face of U1 directly and cameras C2 and C3 record the scenery. The figure is adapted from [112]

in various contexts, since human-human as well as human-machine interaction is present in the dataset. Therefore, it is well suited to serve as a reference dataset comprising a multitude of possible situations and user states, as hardly any bias is introduced due to the recording setup and experimental design. For an exact and thorough usability analysis please refer to [113–115].

Each dialog involves two human participants, who interact with the system S operated by the wizard. The system reacts to questions or gives hints about possible restrictions or search queries. The software setup of the system is explained in more detail in [98]. In Fig. 2, the utilized setup of the system is shown.

Participants The participants ($n = 74$) recorded in 37 dialogs were mostly students and employees of the University of Ulm, who gave written consent to participate in this study. They were between 19 and 51 years of age (on average 24.4 years); 31 female, i.e. 41.8 %). The dialogs were recorded in German, the native language of all participants. The shortest recorded dialog lasts 2:43 minutes, the longest lasted 18:24 minutes. For an exact distribution of dialogs and dialog duration please refer to [111].

Audio data The audio data was recorded using three microphones: One lapel microphone for each participant and a room microphone to record the entire scene including the system output. The audio data was recorded at 16 kHz with 16 bit resolution. The recordings of the high quality AKG CK 97-C lapel microphones was transmitted using the AKG PT40 transmitter. External Creative USB SoundBlaster sound cards were used for the recordings. The AKG C 1000S room microphone was attached to a Sony MZ-R700 Mini Disk recorder.

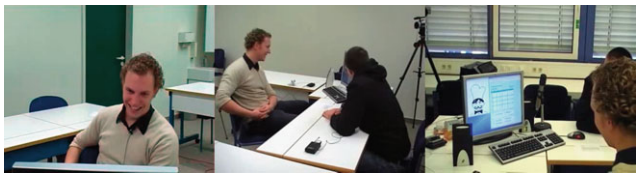

Fig. 3 Typical scene taken from a recording from all three different camera angles. (From left to right: C1, C3, and C2)

Video data The video data was recorded using three distinct cameras as shown schematically in Fig. 2: two cameras are recording the scenery of the interaction, whereas one camera is directly recording the face of the primary user U1. The video of the scenery cameras was recorded at a resolution of 720×576 pixels with a sample frequency of 25 Hz using three JVC GR-D270E cameras. In Fig. 3 a typical scene of the interaction is shown.

3.1 Annotation scheme

One of the challenges, dealing with unscripted and naturalistic interactions, as available in the PIT corpus, is the lack of knowledge about the actual ground truth of the participant's affective states. In contrast to acted emotional data it is not possible to fully control the behavior. On the other side, this lack of control provides naturalistic behavior of users while interacting with machines.

The available labels, developed in this work, for the naturalistic interaction data are shown in Table 1. This layered set of labels was iteratively developed in discussions with the annotators. It allows for comparatively good agreement between annotators, and is motivated as a subset of the literature (compare Sect. 2.2); it is extendable to individual necessities. The annotations are provided in independent layers: subject state, talk style, events, focus of user, and dominant dialog role (i.e. one of the interlocutors frequently holds the turn and speaks for a large part of the interval). Further, alongside these layers, the focus of attention and the actual gaze direction of the primary subject was annotated and analyzed in [115].[11] The layers are organized with respect to criteria of complexity: the individual layers contain groups of labels that have similar time resolutions; the layers events and focus can contain short annotations, whereas the dominant dialog role and subject states can contain more stretched annotations. Further, the different layers require the fusion of different sources or sensors; talk style is mainly inferable from audio and speech, whereas the gaze is solely dependent on the video; the layers dominant dialog role and subject state further require the integration of context. Similar, layers or fusions could be envis-

[11] Unfortunately, the camera setup only allows the annotation of the gaze of user U1, as no frontal view of U2 is available.

Table 1 Extended list of label groups and organization in layers as introduced in [99]. The top and most abstract level is the subject state layer. Lower levels are more objective observations and comprise the social signals

Level	Label	Meaning
Subject state	Interested	Listening (not active), showing interest, reading (silent/loud)
	Uninterested	Distracted, uninterested, not paying attention
	Surprised reacting	surprised, facial expression, utterance of surprise
	Embarrassment	Embarrassment, insecure, blushing, confused
	Impatient	Commenting waiting, impatient movement
	Stressed	Seeming stressed (work, appointments), hasty behavior
	Negative accepting	May be compromising, disappointment
	Positive accepting	Pleased with outcome, acceptance, positive perception of offer
	Disagreement	Disagreeing with the outcome but not accepting yet
Talk style	Commanding	Non-natural command style talk, imperative speech
	Off-Talk	Non-related to topic or HCI
	Ironic	Speaking ironically about something
	Explaining	Pedagogical, arguing, giving facts
	Active listening	Nodding, back-channeling, nonverbal communication
	Question	Posing a question go get information
	Thinking	Loud thinking, pausing, "hm… what shall we do?"
	Reading	Menu/map is presented on the screen; user U1 reads out to user U2
Events	Laughs	Loud laughters, silent ones, prominent smiles
	Silence	Agreeing or disagreeing silences
	Exciting moments	For the participants
	Topic shifts	Change of topic
	Waiting	Waiting for a reply (mostly due to wizard of Oz lags)
Focus of user U1	User U2	The focus lies on person B
	System	The focus lies on the system
	Others	The focus is something else
	Changing focus	There is a shift of focus (i.e. phase of head or eye movement)
Dominant dialog role	User U1/U2	One person is dominant (longer periods)
	Equally active	Lively conversation, back and forth between participants
	Equally passive	Slow and boring conversation

aged in a final setup of a layered architecture as shown in Fig. 1.

In all layers the annotations are temporal attributes and can be assigned with varying lengths and offsets, which renders the evaluation of inter-rater agreement difficult.[12] This temporal flexibility allows the annotator to describe the data in four independent time scales and with additional information, which is according to [20] an important requirement for dealing with emotions or dispositions overall.

Using this annotation scheme, 15 out of the previously mentioned 37 dialogs were annotated using the labeling tool ANVIL [52]. In order not to introduce any bias, the annotators had to annotate the subject state layers of the dialogs in first screenings before annotating the other layers. Each lower level was then annotated separately in consecutive annotation runs.

The distribution of labels over the dialogs and subject roles is listed in Table 3 in Appendix. Additionally, the average durations and the standard deviations are listed there; all of the categories have occurred in the dialogs and significant differences between the two human dialog partners U1 and U2 are revealed. To infer relevant behavioral cues, e.g. talk styles, and events, for the subjective annotation of the subject states, we further analyzed the results of the annotations, as explained in Sect. 3.2.

[12] Standard measurements such as Cohen's κ are designed for atomic entities or pre-segmented samples of the data, which are not available [19].

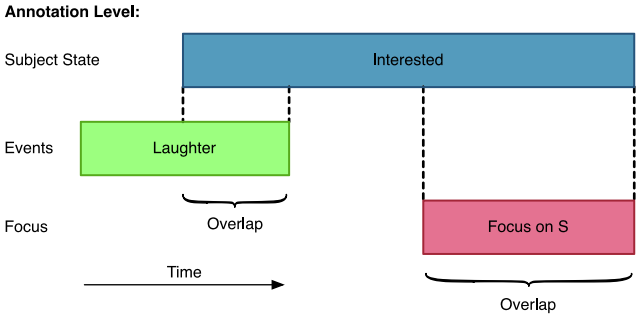

Fig. 4 Illustration of the overlap calculation relative to the overall length of the subject state annotation (here interested), with the lower level observations (e.g. focus and laughter)

3.2 Evaluation and discussion

In the following three labels, namely interested, positive accepting,[13] and negative labels[14] are statistically analyzed in order to find significant correlates between annotations.

In order to measure the coincidence of the annotations the subject state with the behavioral cues of the other layers, the relative overlap of these labels with the subject state was measured for all the annotated dialogs. The relative overlap r is calculated as the overlapping length o of the lower objective label with respect to the length of the subject state annotation l: $r = \frac{o}{l} \in [0, 1]$, illustrated in Fig. 4. The result is evaluated using box plots where brackets with * or ** indicate significant ($p < 0.05$) and highly significant ($p < 0.01$) differences in the overlaps calculated using paired t-tests. The boxes denote 50 % of the data and the median value is shown as the middle line of the plot. Whiskers include 1.5 times the standard deviation of the data and outliers marked as crosses are further away from the median.

In Fig. 5, it is seen that the focus of attention towards the system differs significantly over the three targeted subject states. In detail, U1 is labeled significantly ($p = 0.003$, see ①) more as focusing on the system while he is labeled as interested contrary to negative labels. Further, he significantly ($p = 0.031$, see ②) focuses the system more while interested in contrast to the label positive accepting.

In Fig. 6, the dominant dialog role annotations are compared to U1's subject states: if the system takes over the dominant role in the conversation, highly significant support for the state interested is found (vs. positive accepting $p = 0.001$, see ①; vs. negative labels $p < 0.001$, see ②). It is also seen that if all participants are equally active in the dialog the state positive accepting is significantly ($p = 0.043$, see ③) overlapped to a higher extent.

[13] An offer or suggestion of the system is perceived positively by the subject.

[14] All negative subject states combined, i.e. uninterested, embarrassed, impatient, stressed, negative accepting, disagreement.

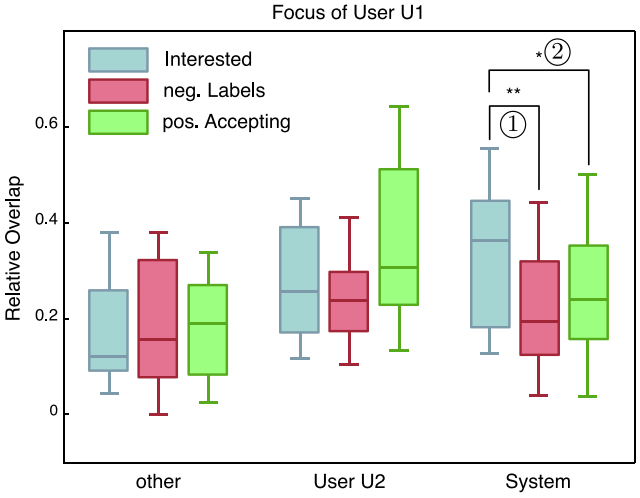

Fig. 5 Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of user U1 with U1's focus (eye gaze) towards the system, user U2, or elsewhere (other)

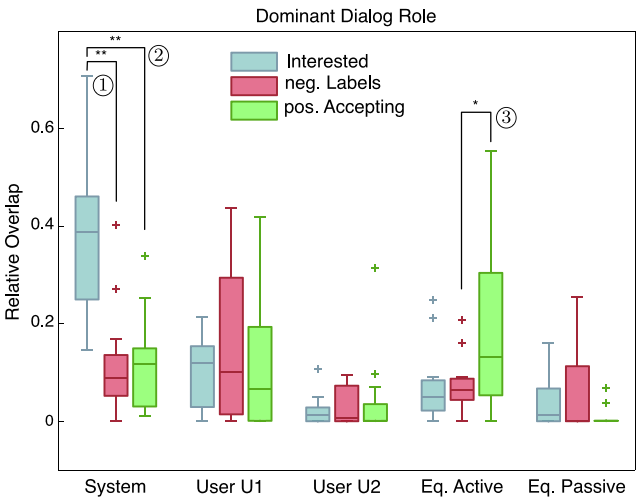

Fig. 6 Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the dominant dialog role annotations

Single turns, utterances or paralinguistic cues of subject U1 are evaluated in Fig. 7. It is seen that the overlap for active listening, including many feedback and backchanneling utterances and paralinguistic cues, such as "um", or "hm", is highly significantly larger for the state of interested as opposed to the other two categories (vs. positive accepting $p = 0.001$, see ①; vs. negative labels $p < 0.001$, see ②). Further, the amount of overlap with respect to the talk style explaining for positive accepting is significantly higher compared to interested ($p = 0.028$, see ③). The overlap of questions posed is not significantly higher if the user is interested, whereas the annotations of commanding are significantly overlapping more with the negative labels as with

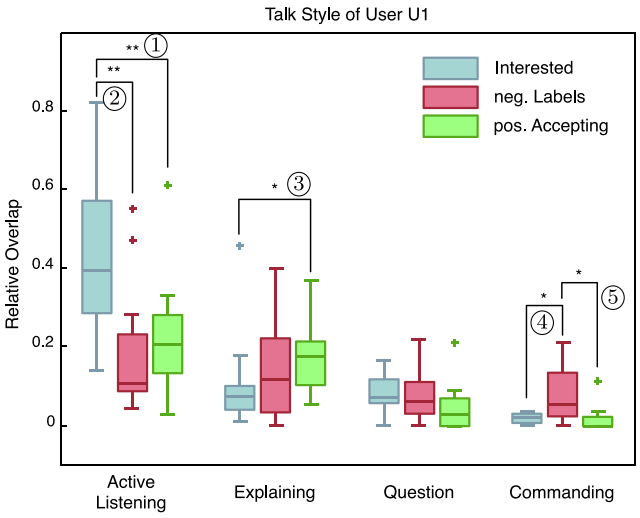

Fig. 7 Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the talk style/utterances of user U1

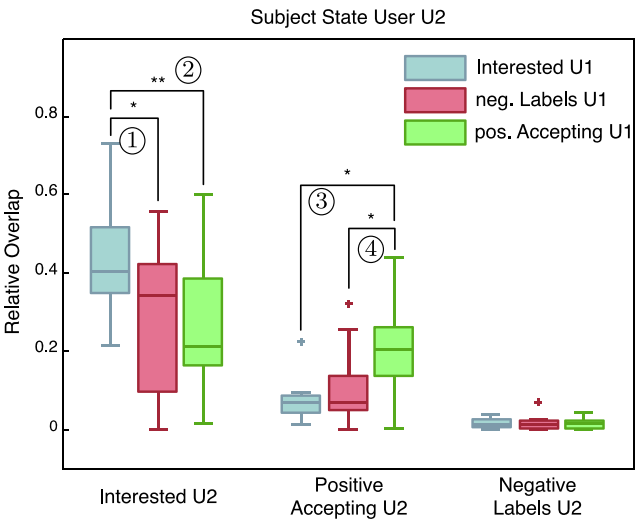

Fig. 9 Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 and U2

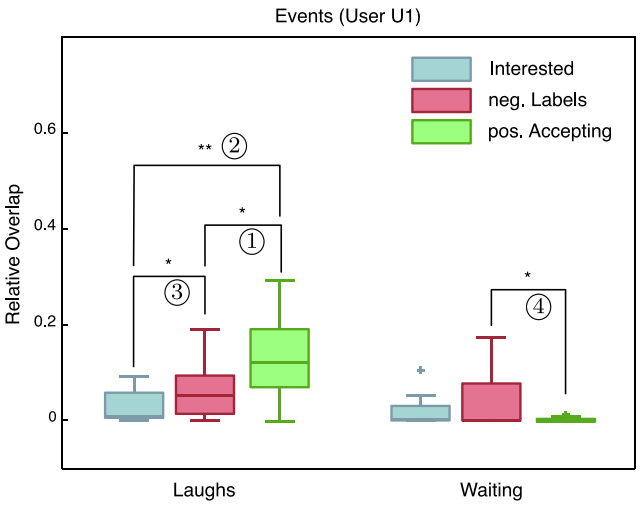

Fig. 8 Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the two most frequent U1 related events (laughs and waiting)

the two other categories (vs. interested $p = 0.012$, see ④; vs. positive accepting $p = 0.019$, see ⑤).

Additionally, Fig. 8 shows the relevance of the labels subsumed in "events" for the identification of the user's state in the interaction. Laughter overlaps significantly more in the state of positive accepting as opposed to the negative annotations ($p = 0.012$, see ①) and highly significantly more in contrary to interested ($p < 0.001$, see ②). This finding supports the idea that U1 is commenting positively perceived suggestions with a surprised or pleased smile or laughter. Further, the overlap of laughter with the negative labels in comparison to interested is as well significant ($p = 0.018$, see ③). Figure 8 also shows that the relative amount of overlap of the annotation of waiting with negative subject state labels is significantly higher than the one for positive accepting ($p = 0.046$, see ④).

In Fig. 9, a comparison of the subject state of U1 to the one of U2 is shown. The labels interested and positive accepting correlate individually between both users forming some sort of common interactional state: The relative overlap of the interested state of U2 with the state of interested of U1 is significantly higher ($p = 0.022$, see ①) than the negative label overlap and highly significant ($p = 0.001$, see ②) for positive accepting. This indicates that U1 and U2 often share the same state of interest.

Further, if U1 is in the state of positive accepting we see significantly higher overlaps for the subject state positive accepting of U2 (vs. interested: $p = 0.017$, see ③; vs. negative labels $p = 0.010$, see ④), indicating that U1 and U2 often share the state of positive accepting. This in turn shows that some of the states annotated in this study are actually shared between the interlocutors and refer to an overall conversational state rather than to an internal or personal state, which was hypothesized in the introduction of this paper.

Overall, the annotation conducted for the PIT corpus reveals several behavioral cues, that indeed coincide significantly more often with certain subject states than with others. These results, together with others found in the literature e.g. [8, 123], motivated us to develop an architecture capable of the spatiotemporal integration of behavioral cues in order to reveal hidden communicational states or social signals. The following Sect. 4 reports some of our previous work to detect, track, classify and fuse behavioral cues, that were developed in parallel to the analysis and annotation of the PIT corpus. As the annotation process of the PIT corpus was extremely time consuming, we developed the approaches on several available different and sometimes more restrictive datasets allowing the preliminary analysis of our machine

learning approaches. The PIT corpus itself is a rich multimodal HCI corpus, that is due to its complexity and the unconstraint recording extremely difficult to process. The PIT corpus itself is a multimodal HCI corpus which provides a rich source of a multitude of communicative features in the auditory and visual channels. Due to the complexity of the corpus with the sequences of unconstrained communication also sets a high demand to its evaluation and annotation. Consequently, any labeling requires focussing of resources and a prelude of detailed planning, e.g., for considering the communicative gestures at the targeted focus of the investigation.

4 Technical approaches

4.1 Generic layered architecture

As shown in the previous sections, the communication state can be inferred from the integration of behavioral cues expressed by the interlocutors, therefore, the proposed architecture processes data on different layers in order to be able to extract temporal patterns of the observed behavior. For simplicity we consider three layers of data streams, without loss of generality. The first layer directly processes the features $\tilde{\mathbf{X}}$ being extracted from the various modalities (i.e. audio, video, etc.). The second layer handles the stream of classification results $\tilde{\mathbf{Y}}$ representing the low level observations, i.e. nonverbal and paralinguistic behavior. The last stream corresponds to the output stream $\tilde{\mathbf{Z}}$, representing the user's states or communicational dispositions we aim at. Note, that theoretically it is possible to train further layers on top of that three layer architecture or to insert intermediate layers.

The data stream of the second and third layer is obtained by shifting a window on the underlying layer and feed the selected input into the corresponding classifiers. We denote the windows on the first and second layer by $\tilde{\mathbf{X}}_{t_1} = (\tilde{\mathbf{x}}_{t_1} \dots \tilde{\mathbf{x}}_{t_1+\hat{T}_1})$ and $\tilde{\mathbf{Y}}_{t_2} = (\tilde{\mathbf{y}}_{t_2} \dots \tilde{\mathbf{y}}_{t_2+\hat{T}_2})$.

A schematic illustration of the framework is shown in Fig. 10. The figure shows that, the higher layers integrate information of coarser time granularities. The last observation of the second layer $\tilde{\mathbf{y}}_{t_2+\hat{T}_2}$ is obtained using $\tilde{\mathbf{X}}_{t_1}$ as input. The first entry of the window of the second layer $\tilde{\mathbf{y}}_{t_2}$, however, is obtained by processing $\tilde{\mathbf{X}}_{t_1-\hat{T}_2}$. Therefore, the window $\tilde{\mathbf{Y}}_{t_2}$ integrates information of the first layer starting from $t_1 - \hat{T}_2$ and ending at $t_1 + \hat{T}_1$. The data stream of the second and third layer is obtained by concatenating the classification results of the lower layers.

4.2 Conditioned hidden Markov model

The conditioned hidden Markov model (CHMM) extends the classical HMM by the assumption that the hidden states

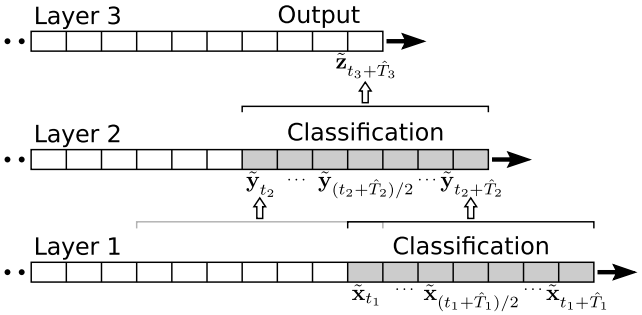

Fig. 10 Exemplary layered arrangement of the hierarchical architecture processing three layers of data streams. Layer 1 processes the features $\tilde{\mathbf{X}}$ which are directly extracted from the multiple sensors. Layer 2 handles the stream of classification results $\tilde{\mathbf{Y}}$, i.e. the social signals found in the features. Finally, Layer 3 corresponds to the output layer representing the user's state or communicational disposition. It is seen that the time granularity from layer to layer increases, as one box in Layer 3 corresponds to $\hat{T}_2$ boxes of Layer 2 and one box of Layer 2 corresponds to $\hat{T}_1$ boxes of Layer 1 respectively

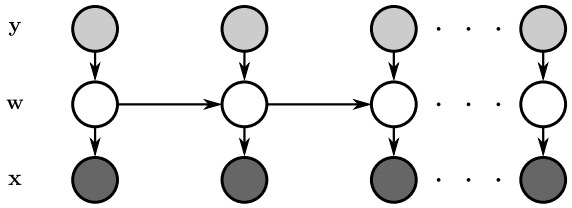

Fig. 11 Graphical model of the conditioned hidden Markov model. The *gray nodes* are required for learning the parameters of the model but only the *dark gray nodes* will be used testing a new sequence

are influenced by an underlying cause, namely the class to be detected. As a result the hidden states can be shared by different classes, which is plausible as actions or communication states may have an overlap in the observation space and can only be differentiated by observing the hole sequence.

The Markov chain of the CHMM is shown in Fig. 11 and consists of a sequence of hidden random variables $\mathbf{w}$, influencing the observations $\mathbf{X}$. The hidden variables themselves are influenced by a sequence of random variables $\mathbf{y}$, modeling causes that induce the selection of the hidden random variables.

The joint probability for a given sequence of observation and hidden states and labels is evaluated by

$$\begin{aligned} p(\mathbf{X}, \mathbf{w}|\mathbf{y}, \lambda) = & p(\mathsf{w}_1 = w_1|\mathsf{y}_1 = y_1, \boldsymbol{\pi}) \\ & \cdot \prod_{t=1}^{T} p(\mathbf{x}_t = \mathbf{x}_t|\mathsf{w}_t = w_t, \theta), \\ & \cdot \prod_{t=2}^{T} p(\mathsf{w}_t = w_t|\mathsf{w}_{t-1} = w_{t-1}, \mathsf{y}_t = y_t, \mathbf{A}) \end{aligned} \tag{1}$$

where the time steps are indexed by t and $\lambda = \{\boldsymbol{\pi}, \mathbf{A}, \theta\}$ the set of parameters of the CHMM. The probability for the initial hidden state is given by $\boldsymbol{\pi} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{y}|}$ while the probability of the transition between the hidden state is given by $\mathbf{A} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{w}|\times|\mathsf{y}|}$. The parameters describing the emissions are given by θ and can model any observation like the classic HMM (e.g. normal distributions, Gaussian mixture models, or Bernoulli distributions).

The expectation-maximization (EM) algorithm is maximizing the expected complete data log-likelihood given by

$$\begin{aligned} Q\big(\lambda, \lambda^{\text{old}}\big) &= \sum_{\mathbf{w}\in\mathbf{w}} p\big(\mathbf{w}|\mathbf{X}, \mathbf{y}, \lambda^{\text{old}}\big) \ln p(\mathbf{X}, \mathbf{w}|\mathbf{y}, \lambda) \\ &= \sum_{j\in\mathsf{w}_1} \gamma_{1,\mathbf{y}}(j) \ln \pi_{j,y_1} \\ &\quad + \sum_{t=2}^{T} \sum_{i\in\mathsf{w}_{t-1}} \sum_{j\in\mathsf{w}_t} \xi_{t-1,t,\mathbf{y}}(i,j) \ln A_{i,j,y_t} \\ &\quad + \sum_{t=1}^{T} \sum_{j\in\mathsf{w}_t} \gamma_{t,\mathbf{y}}(j) \ln p(\mathbf{x}_t|\mathsf{w}_t = j, \theta), \end{aligned} \tag{2}$$

where the belief for being in state i and having the label y_t at time t is given by $\gamma_{t\mathbf{y}}(j)$, and the probability for being in state i and pass over to state j and having the label y_t at the time $t-1$ and t is given by $\xi_{t-1,t,\mathbf{y}}(i,j)$. For a given label sequence, the probabilities can be computed recursively using the forward and backward variables

$$\begin{aligned} \alpha_{t,\mathbf{y}}(j) &= p(\mathbf{x}_t|\mathsf{w}_t = j) \\ &\quad \cdot \sum_{i\in\mathsf{w}_{t-1}} \big(\alpha_{t-1,\mathbf{y}}(i)\ p(\mathsf{w}_t = j|\mathsf{w}_{t-1} = i, y_t)\big) \quad \text{and} \\ \beta_{t,\mathbf{y}}(j) &= \sum_{i\in\mathsf{w}_{t+1}} \big(p(\mathbf{x}_{t+1}|\mathsf{w}_{t+1} = i) \\ &\quad \cdot \beta_{t+1,\mathbf{y}}(i)\, p(\mathsf{w}_{t+1} = i|\mathsf{w}_t = j, y_{t+1})\big), \end{aligned} \tag{3}$$

with the starting conditions

$$\begin{aligned} &\alpha_{1,\mathbf{y}}(j) = p(\mathbf{x}_1|\mathsf{w}_1 = j)\, p(\mathsf{w}_1 = j|y_1) \quad \text{and} \\ &\beta_{T,\mathbf{y}}(j) = 1. \end{aligned} \tag{4}$$

The probabilities are then obtained by

$$\begin{aligned} \gamma_{t,\mathbf{y}}(j) &= \frac{\alpha_{t,\mathbf{y}}(j)\beta_{t,\mathbf{y}}(j)}{p(\mathbf{X})} \quad \text{and} \\ \xi_{t-1,t,\mathbf{y}}(i,j) &= \frac{\alpha_{t-1,\mathbf{y}}(i)\beta_{t,\mathbf{y}}(j)}{p(\mathbf{X})} \\ &\quad \cdot p(\mathbf{x}_t|\mathsf{w}_t = j)\, p(\mathsf{w}_t = j|\mathsf{w}_{t-1} = i, y_t). \end{aligned} \tag{5}$$

As a result, the conditioned probability to observe a sequence given a label sequence is given by

$$p(\mathbf{X}|\mathbf{y}) = \sum_{i\in\mathsf{w}_T} \alpha_{T,\mathbf{y}}(i) \tag{6}$$

and hence the joint probability is given by $p(\mathbf{X}, \mathbf{y}) = p(\mathbf{X}|\mathbf{y})p(\mathbf{y})$. The distribution on the class labels is then obtained by

$$p(\mathbf{y}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{y})}{\sum_{\mathbf{y}\in\mathbf{y}} p(\mathbf{X}, \mathbf{y})}. \tag{7}$$

To compute the joint probability all combinations of the label sequence $\mathbf{y}$ need to be considered. This procedure is computationally expensive. Since we use a sliding window only slow changes within a sequence can be assumed, and therefore the label sequences evaluated are restricted to only one class.

During maximization we have to deal with the problem of unbalanced data sets (e.g. as explained in Sect. 4.3.1, laughter is only present in about 10 % of the conversation). In case a class holds the majority, the parameters will be dominated by the corresponding distribution. In order to prevent the CHMM to converge into this maximum, the numbers of samples per class need to be adjusted by having repeated entries in the dataset drawn by chance from the underrepresented class.

Since emissions are now shared by classes, the approach promises to generalize better. However, during training it might occur that hidden states are never visited by a class and the corresponding transition probability converges to zero. Therefore, it is advisable to replace these values by a numerical quantity close to zero. The most important difference to classical HMM is the funded distribution over classes which holds more information than the discrete class assignment itself.

4.3 Mechanisms for analyzing non-verbal cues

In this section we highlight some components of the suggested framework and architecture along with some exemplary results. In particular, laughter spotting, as motivated above, is investigated from the view-point of online classification in natural conversations. Furthermore, we show results of investigating the classification of voice quality as a paralinguistic quality for affective user-state and additional categorical information estimation. As briefly mentioned in the motivation to this work, the visual channel—and different features and combinations thereof—convey important signals of non-verbal communication, such as, attending/averting, shifting focus of attention, or pointing and gesturing through changes in head and body pose or movement

signatures of the limbs and hands. We exemplary show results of estimating the head pose as an indicator of the interlocutor's direction of his/her focus of attention. Also, we show an example of robustly extracting the arms and hands as a precursor for estimating temporal signatures of hand gestures.

In addition, results of fusing auditory features with results from the visual processing channels are briefly demonstrated. Finally, we also demonstrate results that were gathered at the higher level of activity recognition in an interaction scenario. Thus, these demonstrate how the proposed architectural framework has already been pushed forward to spotlight on an emerging set of functionalities that operate on different layers.

4.3.1 *Audio-based behavioral cue detection: laughter spotting and voice quality estimation*

Spotting laughter in naturalistic conversations In [93, 96] we investigated the capability of sequential classifiers to spot laughter in natural multiparty conversations, described in [16]. The dialogs examined have been recorded using unobtrusive recording devices, namely one centrally placed microphone and a 360 degree camera. Despite the simple setting and high variability of the recorded utterances and laughs, high accuracies in spotting laughter could be achieved. Laughter is arguably one of the most important and universal paralinguistic behavioral cues [50, 53, 61, 120]. It is crucial for the detection of highly engaging points in conversations and is a direct indicator for the positive quality of interaction as mentioned in [61] and shown in the annotation analysis in Sect. 3.2.

Laughter appears in a large variety of forms and types (e.g. inhaled, exhaled, snort like laughs as well as laughter bearing various meanings, e.g. humorous, nervous, or social laughter [12]). A thorough analysis of the laughter's acoustic features in [3, 5] revealed interesting fundamental frequency (f_0) characteristics of single laughter calls.[15] The so called f_0 excursions and change[16] within calls were quite significant (e.g. around 50 Hz excursion and 45 Hz change for male laughers, and 80 Hz excursion and 60 Hz change for females); features capturing these variations, such as modulation spectral features (ModSpec) could have quite significant impact on the detection performance [43, 67].

Apart from our studies, a considerable amount of effort has been invested into the analysis of automatic uni- and multimodal laughter detection systems in the past few years [5, 50, 81, 87, 120]. However, most of the studies did not suffice all of the criteria required for the direct applicability of the approaches in online laughter spotting scenarios. Either, manual initialization is required, the performance is evaluated and trained on manually segmented laughter episodes, or sterile and controlled laboratory recordings are evaluated.

Therefore, we analyzed natural multiparty conversations with methods sufficing the mentioned requirements in Sect. 2.4 in [93, 96]. In detail, we extract three independent feature streams from the audiovisual data, namely ModSpec features, and perceptual linear prediction (PLP) coefficients [42] from the audio stream, and coarse movement related features from the video stream [28]. Further, these features were utilized in several multimodal experiments using sequence classifiers, such as HMM and ESN processing multiple sources, and decision level fusion.

With both approaches, we achieved high detection accuracies (6.5 % error for the best HMM approach and 9.1 % error for the ESN). However, error rates do not tell the whole story, as the datasets used for spotting rare events, such as laughter in conversations, are often skewed with respect to the amount of background data versus the target (in the analyzed corpus the ratio is about 10:1). Therefore, we calculated the F_1 measure[17] ($F_1 = 0.72$ for the HMM, with 0.8 recall and 0.64 precision; $F_1 = 0.63$ for the ESN with 0.81 recall and 0.52 precision) in order to ensure a fair comparison of performances.

In conclusion, it is possible to spot important paralinguistic behavioral cues, such as laughter, from unobtrusive audiovisual sources, with good accuracies in an online manner using multimodal sequence classifiers. However, one of the conclusions in [93] is, that even though multimodal analysis often improves the performance it is not a straight forward process. Further, some of the laughs could not be detected with any of the proposed setups and configurations, leading to the assumption that there is still space for improvement (e.g. other purpose tailored features, more complex decision integration, etc.).

Classifying voice quality using fuzzy-input fuzzy-output support vector machines In [94] the classification of the paralinguistic phenomenon of voice quality is investigated, as it is important for manifold applications and is a strong indicator for the speaker's attitude, affective state and mood. Voice qualities, such as breathy, creaky, whispery, tense, and harsh voices, convey important information about the affective state, as well as gender, and age of the speaker [39, 65, 69, 132]. In [39] and [40] it is even stated that voice

[15] A laugh bout contains multiple calls (e.g. the typical repetition of 'ha').

[16] Excursion is defined as the difference between the maximum and minimum f_0 within a call, and the change is defined as the absolute value of the difference between the f_0 at the call onset and the one at the call offset.

[17] $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$, where P denotes the precision (ratio of hits to all hits and false alarms) and R the recall (ratio of hits to all laughs in the set) of the approach.

quality does in fact contain more information on the affective state of the speaker than intonation.

The term voice quality refers to the timbre or coloring of a speaker's voice. For an individual speaker their voice quality is composed of longer term settings of the vocal system combined with dynamic shifts in the system for communicative purposes [62, 66]. The perception tests conducted for the classification experiments in [94] revealed that the investigated voice qualities breathy, modal and tense, are a rather continuous or dimensional concept than crisp categorical entities. Human experts for example confused breathy and tense samples mostly as being modal, but breathy and tense were hardly ever confused with each other, resulting in an inter-rater agreement of $\kappa = 0.526$ [19].

In order to investigate the usefulness of uncertain and fuzzy information provided by the human experts, we analyzed the classification performance of fuzzy-input fuzzy-output support vector machines (F^2SVM) [119]. These F^2SVM outperformed naive Bayes classifiers and standard SVM with the same configuration in multiple experiments (i.e. 10 fold cross validation, leave one speaker out, and cross corpus experiments) significantly, by solely utilizing the information provided by the fuzzy annotations of the human experts during training on a subset of the voice quality data for which the majority vote of the human annotators always coincided with the actual target label. The inter-rater agreement for this subset is $\kappa = 0.717$. The F^2SVM classified the voice quality samples with an error rate of 12.14 % ($\sigma = 3.11$) in the 10 fold cross validation, 13.88 % ($\sigma = 3.89$) in the leave one speaker out experiment, and 17.66 % in the cross corpus experiment in all cases these error rates are significantly lower as for both reference approaches.

In an additional fuzzy evaluation experiment, we did not treat the output of the F^2SVM as probabilities for each of the three voice qualities and did not choose the most likely one. We allowed mixed states and regarded the fuzzy output as a membership distribution of the sample to all three classes. For this experiment we utilize the D_1 distance measure to evaluate the fuzzy output. D_1 is defined as:

$$D_1(x, y) = 1 - S_1(x, y) \in [0, 1], \tag{8}$$

$$\text{where } S_1(x, y) = \frac{\sum_{i=1}^{L} \min(x_i, y_i)}{\sum_{i=1}^{L} \max(x_i, y_i)} \in [0, 1]. \tag{9}$$

The measure captures the similarity between two fuzzy labels x and y, where L denotes the number of classes. The more similar the target y and the classifier result x are, the closer the value $S_1(x, y)$ will be to 1. The S_1 similarity measure is commonly used for fuzzy classifier fusion using decision templates [56, 58].

This approach resulted in an average D_1 of 0.3884 ($\sigma = 0.0288$) for the cross validation experiments and 0.3877 ($\sigma = 0.0260$) for the leave one speaker out experiments with no statistically significant difference between the automatic approach and the human baseline performance.

The conclusions to be drawn from the analysis and experiments in [94] are: the usage of fuzzy and uncertain information does indeed improve classification results and outperforms standard approaches significantly in certain cases. Further, mixed categories for dimensional and continuous concepts, such as voice qualities and affective states, should be processed in a holistic and generic architecture analyzing social signals on various levels. However, in order to be able to evaluate these results, measures such as the D_1 distance measure need to be established and utilized.

4.3.2 *Visual detection of head pose, limbs, and articulatory activities*

Another important modality for non-verbal communication is given by the visual channel. Visual signals comprise rich behavioral cues to establish sensitivity to emotion and disposition in vision and speech, gesture, touch, and individuality concerning autonomy and personality. In communication processes nonverbal social signaling conveys communicative initiative, determination, interest, attention, relatedness, empathy, etc. [79, 124]. We briefly summarize our approach to reliably estimate the user's head pose direction, one of the most meaningful cues concerning the user's focus of attention relative to the observer or camera (see [64]). In addition, we demonstrate how parts of the upper torso, such as the limbs (arms, hands) can be extracted and utilized to subsequently estimate temporal signatures for gestured articulations. It should be emphasized that we aim at developing mechanisms that function as integral part of the general framework, which is capable of analyzing social signals from a first-person perspective. This is important as a significant amount of contributions visually analyze communicative signals and behavior from a third-person perspective such as to automatize evaluative function to derive significant features in multi-party interactions [79].

Head pose estimation Vision-based estimation of the human head pose received more and more attention over the last decade [70]. Many of the approaches proposed in the literature still suffer from large errors and redundancies in the feature extraction and matching mechanisms, vulnerability to variations in illumination conditions, or size changes due to variable camera-actor distances. We propose a scheme based on stereoscopic vision that overcomes these difficulties by operating on features of intermediate complexity derived from a biologically inspired hierarchical processing architecture. Using intermediate-level features increases the robustness of the stereo matching, while the false detection rate is reduced. The approach mainly consists of two processing steps, namely (i) the detection and localization of

four facial features (the eyes and the mouth corners) within the images of a stereo pair and (ii) the calculation of the disparity values between each pair of corresponding features. For the detection of the head and the facial features, we modified the biologically inspired object-recognition model proposed by Mutch and Lowe [71]. The processing proceeds hierarchically through a sequence of alternating layers of local linear filtering operations to combine input features of increasing complexity, and local non-linear pooling of the input over a neighborhood to achieve invariance against variations in position, size and rotation. We employ five stages of processing. First, the original image is transformed into a pyramidal representation of different spatial scales. Second, each scale is convolved with 2D Gabor filters of four orientations resulting in a 4D feature representation of position, scale and orientation. This approach is performance optimized by using fixed filter sizes at the different pyramid levels (unlike, e.g., [108]). In the third stage, filtering results are pooled over the same orientations within a local neighborhood (for details see [64]). In the fourth stage, a template matching operation is utilized using patches from the last pooling stages and a number of learned prototypes. Briefly speaking, the learning algorithm selects the most descriptive and discriminative prototypes among an exhaustive number of patches randomly sampled during the learning process. Disregarding the learning process, the prototypes themselves are filters representing features of a higher complexity. In the last stage, the responses of prototypes from the fourth stage are pooled over all positions and scales by choosing the maximum value for each prototype, yielding a single feature vector. These vectors serve as input to a linear SVM, used for the classification of the facial features. After the successful localization of the intermediate level facial features in the stereo image pair, the associated disparities are determined by maximizing the correlation of a feature in the left image and its counterpart in the right image within a local neighborhood. It is important to stress, that we do not need to estimate a dense disparity map and thus considerably reduce the computational costs. Given the disparity values as well as the focal length and the baseline of the stereo camera system, the world coordinates of the facial features can easily be derived. We have reasoned, that the orientation of the head can be estimated by fitting a plane on four facial feature positions located in space. These features are the centers of the left and right eye and the two corners of the mouth which together approximately span a plan (the so-called facial plane). After the disparities of the matching features have been determined, their 3D depth values are calculated. Finally a regression plane is fitted to the four points such that its orientation can be used as an estimate of the head pose. We used artificial input images with known ground truth, as well as images achieved under almost unrestricted real word conditions to evaluate the precision of the proposed approach (see Fig. 12).

Limb detection and estimating articulatory activity We suggest to employ a task-driven processing mechanism using visual routines [107, 121] in order to selectively constrain the search space for analyzing possible shape configurations. Initial contrast detection is accomplished by using the processing cascade as outlined above. Oriented contrasts can be subsequently grouped to form extended boundary signals [127] followed by the estimation of a symmetry axis. These static features are combined with evidences derived from motion and spatiotemporal occlusion signals. The detection of spatiotemporal occlusion from motion is motivated by investigations of the deletion and accretion patterns of optical texture in the presence of mutual surface occlusions. We employ here an extended mechanism detecting discontinuities in integrated motion patterns. Motion discontinuities are combined with responses from the detection of temporal changes in motion energy encoding occlusion and dis-occlusion regions in the motion field [7]. We employ this motion-based mechanism for segregating the arm limbs and hands from background. Upper and lower arm segments can be compactly described by a symmetry axis. We incorporate an approach proposed by Curio and coworkers to compute medial features from grey level input images [33] (see Fig. 13). In a nutshell, the algorithm utilizes an energy minimization approach based on two sequential stages of processing by estimating a vector field from diffusing boundary gradients, and subsequent detection of local sinks in the resulting vector fields. The orientation of the upper arm limb (at the shoulder joint) is indicative for the potential locations of the lower arm and the hand, based on the degrees of freedom at elbows and wrists. Based on anthropometric constraints as well as the degrees of freedom of the limbs the lower arm and hand is bound to a circular sector of potential image appearance defined by the body and the end of the upper arm axes. The outer limb components can be detected by a further stage of symmetry-based detection as well as an active segmentation stage utilizing the mechanism by [1]. The robust segmentation of the body limbs, e.g. upper and lower arms, hand and even individual fingers convey further information about, e.g., exposure, gestures, self-presentation, and conversational distance [2]. For real scenes, we suggest the sequentially organized extraction of arm and hand poses by an attention-guided search process which proceeds in a coarse-to-fine manner. Segmentation of the hand is triggered by higher-order visual routines which operate to build an incremental representation providing a link to sensory-motor tasks. A target region in the potential sector of occurrence of lower arm limb and hand is identified after the upper arm has been detected. We utilize the approach of [68]. In a nutshell, the algorithm actively centers the local reference coordinate system at a selected target region (simulating an artificial saccadic eye movement). Using a space-variant image representation that centers the high-resolution at the gaze center leads to a simplified segregation

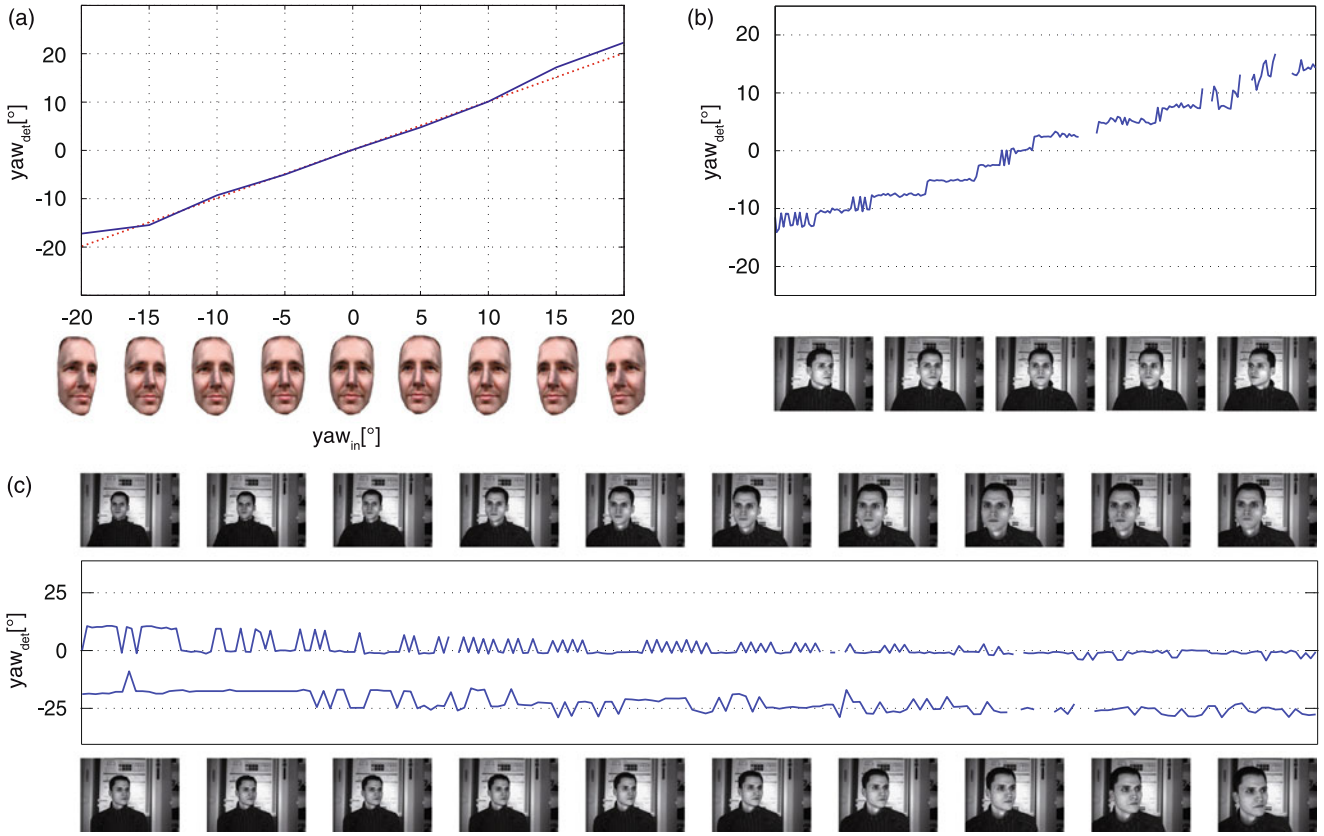

Fig. 12 Evaluation of the stereoscopic head pose estimation. (**a**) Artificially generated stereo images with known ground truth were used to evaluate the precision of the head pose estimation under controlled conditions. For horizontal head poses in a range of $-20°$ to $+20°$ (yaw angle), the error remains below 3° (with larger yaw angles producing larger errors). (**b**) In addition, we tested the proposed head pose estimation approach under real world conditions, using a Bumblebee®2 Stereo Vision Camera System (Point Grey Research Inc.). The subject was instructed to rotate his head systematically from the left to the right. Although no ground truth data is available, it can be seen that the estimated head pose follows the pose characteristic of the head. (**c**) The capability of the proposed approach under varying camera-actor distances was tested using two real world sequences with different but constant yaw angles (above and below the plot). As it can be seen, there is only a little variation in the estimation quality for the head poses at different distances. It is important to note that the apparently large differences of the estimated head poses in the left half of the plot (large distance between camera and actor) are caused by a mere difference of one pixel in the estimated disparity value. Increasing the camera-actor distance more and more decreases the resolution of the face and consequently the number of distinguishable head poses

of the focused target region from the current background (that surrounds the current target). A binary graph cut algorithm segregates the figural segment (the hand) against the peripheral background using the figure boundary to steer the min-cut segmentation. The segmented target region can be tracked over time to estimate characteristic temporal signatures in nonverbal communication. We show a result of the signatures derived from the variance of the Euclidean distance between the hand in pairs of successive frames in a video stream (see Fig. 14). The variance is considered as an indicator of gesticulation activity to emphasize content in verbally communicated message content.

4.3.3 Multimodal and layered fusion experiments

Research activities in facial expression recognition and speech based emotion recognition [77] are usually performed independently from each other. But in almost all practical applications people speak and exhibit facial expressions at the same time, and consequently both modalities should be used in order to perform robust affect recognition.

Therefore, multimodal, and in particularly audio-visual emotion recognition has been emerging in recently [102]; multiple classifier systems have been widely investigated for the classification of human emotions [24, 97, 100, 105, 125].

In the following, we elaborate on two architectures that we investigated recently. The first exemplifies the fusion of audiovisual data for the task of emotion recognition; the other is an exemplary use-case of a layered hidden Markov model for the classification of complex activities.

Audio-visual emotion recognition For the audio-visual emotion challenge (AVEC) [26], we contributed a multi-view architecture for audio-visual affect recognition based

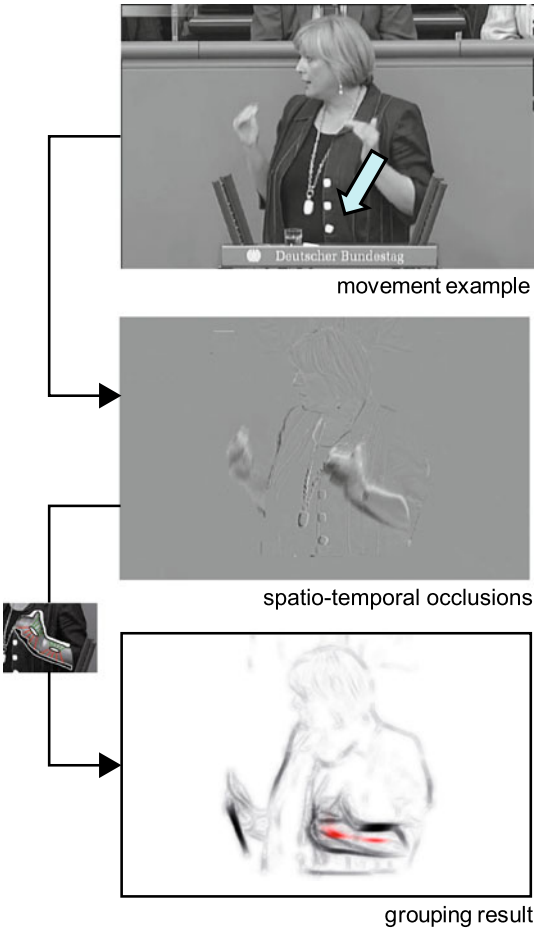

Fig. 13 Detection of the forearm's symmetry axis. Motion-based occlusion and dis-occlusion areas are extracted in an image sequence and used for the calculation of the forearms medial axis

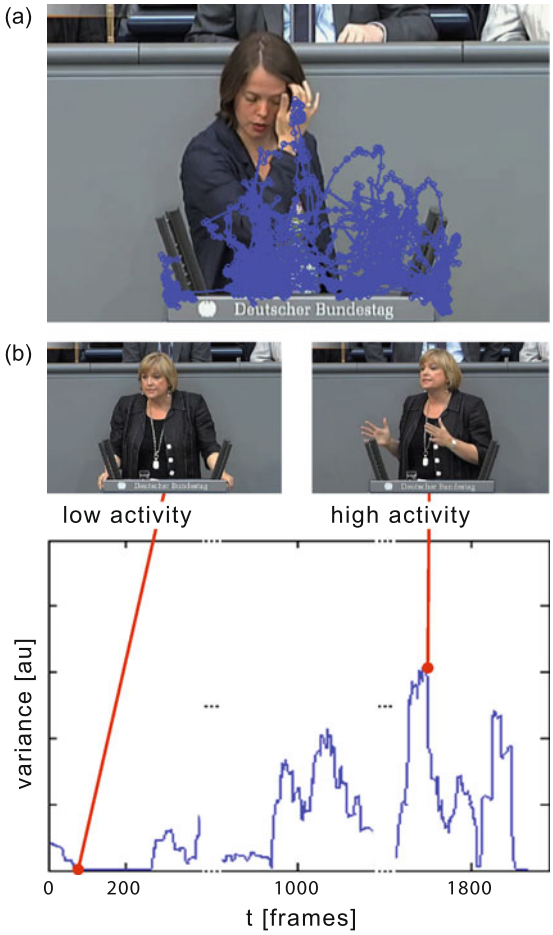

Fig. 14 (**a**) Exemplary temporal signature of a moving hand. (**b**) The variance of the Euclidean distance between one hand in two succeeding frames is used as an indicator of gesticulation activity. One frame for each high, as well as low gesticulation activity is shown

on the principles on multiple classifier systems [37]; the proposed architecture consisted of three parts: an audio-based multiple classifier system trained on Energy, Pitch, RASTA, LPC, and MFCC features, a multiple classifier system trained on form and motion features, and a trainable fusion mapping consisting of multilayer perceptrons (MLP). The task was to accurately predict four affective dimensions: arousal, power, expectancy and valance.

For each audio feature a bag of hidden Markov models (HMM) have been trained [9, 86]. The hidden states and the number of mixture components of the HMM classifiers have been optimized using a parameter search, using the challenge's development dataset; the optimal setup was three hidden states and two mixture components in the Gaussian mixture model (GMM) with full covariance matrices. The overall multi-view audio classifier architecture is schematically shown in Fig. 15 and the unimodal classification rates on word level are stated in Table 2.

The video modality was analyzed using support vector machines (SVM) [101]; concatenated vectors of form and motion features were used as input. These vectors have been used to train a ν-SVM with linear kernel and probabilistic classifier outputs according to Platt [83]. Again a parameter search was applied to obtain optimal parameters for the SVM classifiers. Based on the results of all four label dimensions an intermediate fusion was conducted using an MLP to obtain the final prediction. A scheme illustrating the architecture used is shown in Fig. 16, and classification results on frame level are reported in Table 2.

For the audio-visual fusion the probabilistic outputs of the video stream are collected; subsequently vectors are normalized and decisions are made on word level (as indicated in Fig. 17). By concatenating the results of all label dimensions, a new feature vector is obtained which is then used to train a final fusion layer based on MLP; Fig. 17 shows the overall audio-visual classifier system, while the classification results on word level are given in Table 2; the results are promising and achieved a second and third place in the AVEC competition at ACII 2011 with more than 40 submissions.

Activity classification using layered HMM architecture The layered architecture has already been successfully applied to human activity recognition [35]. The activity recognition setting aims at discriminating a multitude of activities performed by a person (e.g. writing a note, stirring a coffee, drinking coffee, etc.). In a recent experiment the object interacting with person is also detected [36]; the CHMM is then used to incorporate the additional symbolic information into the layered architecture.

These activities can be decomposed into a set of atomic actions (e.g. pick up object, move objects towards head, manipulate object close by the head, lay back object etc.); similar to the behavioral cues for the subject state classification, it is possible to detect the complex activities by integrating over the atomic actions. This behavior is linked to the first challenge in Sect. 2.3. Within the setup the first layer detects the actions and then passes the classification results to the second layer in which the more complex activities are detected. In the proposed HCI scenario the subject's state and communicative dispositions will be decomposed analogously into behavioral cues. However, the relation between them is not as straightforward as for the activity classification task.

The setup's performance was investigated in an offline as well as an online experiment: in the offline experiment actions and activities were pre-segmented for training and testing. Using this data set all examined classifiers achieved error rates of only 2.0 %.

In the online experiment the architecture did not receive the start and end times of the activities; this design fits a real world scenario more appropriately. The data set was not cut into segments and a window is shifted over the continuous data stream (see Sect. 4.1). Additional unseen and untrained background actions (e.g. enter a text into the keyboard) were performed to increase the difficulty. This online experiment clearly shows that investigating pre-segmented data is insufficiently representing the performance of a system in the wild; the error rate was heavily underestimated. While, passing discrete class assignments from the lower layer to the upper reached an error rate of 66.33 %, the utilization of CHMM achieved a significantly reduced error rate of 39.16 % [35]. This result exemplifies the benefit of passing uncertain classification results to the next layer; the integration of uncertain information helps the overall performance to compensate for errors originating in lower levels of the architecture. The utilized CHMM is conform with several of our previously mentioned requirements and challenges. Currently we are working on inferring user attitudes from the stream of detected activities or combined behavioral cues. The architecture is therefore to be extended by a third layer.

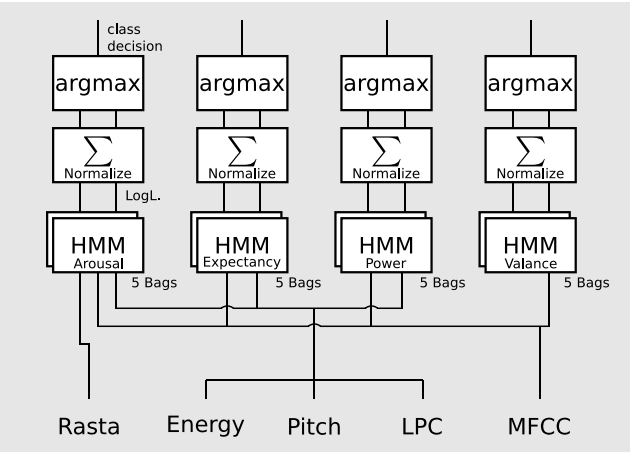

Fig. 15 Audio classifier system. For each label a bag of HMM have been trained on selected features sets

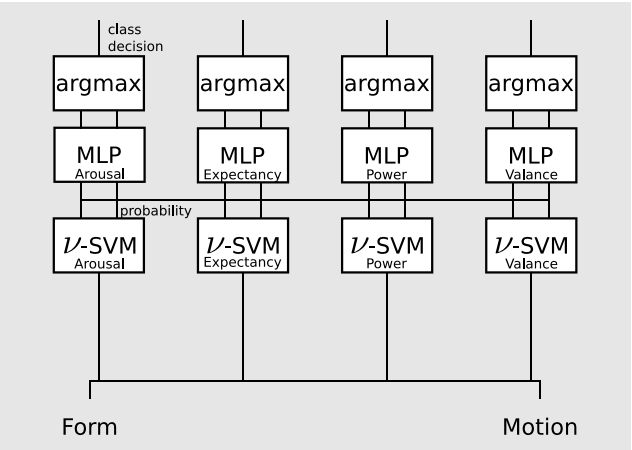

Fig. 16 Video classifier system. The form and motion features are concatenated and used to train ν-SVM for each label dimension. The outputs of the classifiers are used to train a intermediate fusion layer realized by multilayer perceptrons

Table 2 Classification results: The weighted accuracy (WA) corresponds to the correctly detected samples divided by the total number of samples. The unweighted accuracy (UA) is given by the averaged recall of the two classes of a label dimension

	Arousal		Expectancy		Power		Valence	
	WA	UA	WA	UA	WA	UA	WA	UA
Audio	66.9	67.5	62.9	58.5	63.2	58.4	65.7	63.3
Visual	58.2	53.5	53.5	53.2	53.7	53.8	53.2	49.8
Audio/Visual	69.3	70.6	61.7	60.1	61.3	59.1	68.8	66.4

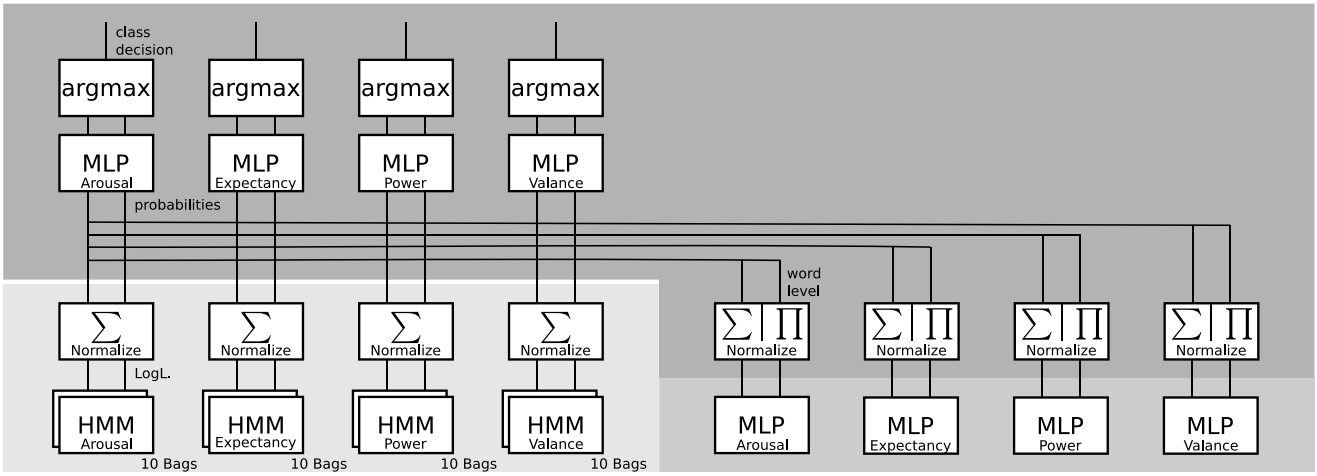

Fig. 17 Audio-visual classifier system. The output of all modalities are collected on word level and used to train a multi layer perceptron for each label dimension

5 General discussion

So far, we discussed a manifold of research fields in this paper. The various aspects that are investigated, and evaluated in the preceding sections are add up and integrated to a bigger picture with our previous work and some related work in the literature.

Relevant HCI user states and suitable annotation schemes ought to be discussed, as the commonly used prototypical emotion categories (e.g. anger, happiness, sadness) rarely occur. Further, annotations are often rigid and limited to fixed utterances [38, 95, 99]. To approach this issue, a hierarchical annotation scheme with labels grouped into several layers is introduced [99]. The correlations between annotations of behavioral cues and user states are analyzed to reveal pattern dynamics that can help reveal user states (e.g. interest, engagement, etc.) [95]. The user states themselves are only indirectly inferable by the annotator using background information (i.e. the context of the interaction) and the observation of the participants' behavior.

The proposed set of communicational dispositions (see Table 1), which is not based on a psychological theory but is rather a pragmatic solution, is well represented in the PIT corpus (see Table 3). The annotations are organized in layers as temporal attributes, and can be assigned with varying lengths and offsets. Hence, it is possible to overlap longer lasting states, such as the dominant dialog role, in one layer and multiple short events, such as the focus of the user, in another layer. This softens the often rigid and strict annotation schemes used in many applications. However, it also increases the complexity of the annotations.

The analysis of the annotations revealed several significant coincidences between the direct observations and hidden user states. For instance, significant correlations of the relative overlap between the user's focus and the state of interest are found (compare Fig. 5); the dialog role and the dynamics of the conversation could also be identified as a relevant cue (see Fig. 6). The dynamics of the conversation are discussed and analyzed in several publications, such as [17, 30, 46, 72]. Additionally, significant coincidences of laughter with the state of positive accepting were found as well, as shown in Fig. 8.

The proposed annotation is, however, not exhaustive and extensions are possible. In [132] for instance it has been shown that voice quality can be used to infer the affective state of the speaker and Sect. 4.3.1 shows the capability to automatically classify it. The subject state layer comprising the so called conversational dispositions is not compulsory, but could in principal be exchanged by other schemes, such as dimensional affect annotations or other categories.

Even though, this annotation scheme is quite advantageous (e.g. flexible and extendable set of categories and layers, relevant subject states, and exchangeable top layer categories), there are also a few non-negligible disadvantages. The time and effort required to annotate in all these different layers for example is large and the issue of how to assess inter-rater agreement of continuous labels is unsolved. The high workload for the annotators may be reduced little by little using bootstrap and partially supervised learning approaches, such as active learning [109] in combination with classification approaches, e.g. laughter spotting or gaze detection (Sect. 4.3).

Active learning can speed up the training of the classifiers as they choose relevant samples[18] themselves in an iterative

[18]Usually relevant samples are those that influence the training the most (i.e. for a support vector machine those samples closest to the separating hyperplane are most relevant for the adaptation of the hyperplane; samples that are far from the hyperplane have hardly any influence and won't help during training).

process. Following this paradigm pre-trained and initialized classifiers can be adapted and personalized iteratively for a particular user, who is frequently interacting with a system.

Within the proposed architecture, the different models can be adapted separately; in particular low level models (i.e. behavioral cue spotters etc.) can be trained independently from the upper layer CHMM rendering the current subject's state [75]. The layered architecture itself is advantageous as the classifiers of the lower layers can be exchanged while the classifier on the upper level may remain unchanged.

Following the requirements for behavioral cue detection (see Sect. 2.4), some successful previous work is shown in Sects. 4.3.1 and 4.3.2. The detection or classification experiments were conducted parallel to the lengthy annotation process of the PIT corpus; in-place and readily available datasets with a subset of the proposed or exclusive (e.g. voice quality) annotations were utilized to test prototype algorithms. Even though, the results were not conducted on the PIT corpus itself but on other more or less constrained datasets, they are representative for the future development of an integrated system.

The study of the layered architecture of HMM using the application of human activity recognition (compare Sect. 4.3.3) showed that uncertainty and problem decomposition are successful strategies. Utilizing classical HMM within this architecture bears the drawback, that solely discrete classes or model likelihoods can be utilized which in general tend to take very small values and therefore are not suitable for training the next layer. To obtain a more qualified probability measure the CHMM as in [35, 36] can be used to render a class distributions for each layer. It could be shown that using the distribution over classes in contrast to discrete class assignments improves the performance of the architecture significantly (see Sect. 4.3.3).

In Fig. 1, a schematic overview of such a multimodal hierarchical architecture is shown. Lower levels are incorporated using sequential classifiers in higher levels. From bottom to top the time windows for decisions are growing, whereas the complexity of the decisions is increasing. Following the requirements and challenges mentioned in this paper, the architecture needs to be able to incorporate the detections and decisions from previous layers; it needs mechanisms to compensate early errors, sensor failures and occlusions. As in [37], data with varying time-constraints (e.g. different sampling rates, expression lengths, or dynamics etc.) need to be integrated and synchronized to a common decision. Further, the increasing time window size from layer to layer implies, that from the amount of data available for training is shrinking in the top layers. Therefore, large amounts of data, rendering the approach expensive in terms of annotation and recording or classifiers with a limited number of trained parameters, are needed.

6 Conclusions and future directions

In this paper we presented an approach towards the analysis of users' states and communicational dispositions in social and human computer interaction. Motivated by the findings in literature and related sciences our approach is based on behavioral cues forming the basic building blocks of the analysis, as they are multimodal indicators for the target states. The temporal and dynamic integration of their occurrences helps to infer the underlying subject state, as shown by the analysis of manually annotated recordings and some previous findings in the literature.

For the automatic inference of the states we propose a hierarchical architecture, based on layered hidden Markov models, capable of achieving this integration, and fulfilling the requirements defined in this paper, such as the capability to deal with uncertain and incomplete data. Supported by findings of our previous work we believe that this step by step approach based on the detection and integration of multimodal behavioral cues can help to achieve the very ambitious goal of user state inference in the future.

Nonetheless, there is a lot of space for further investigations and development; the identification and understanding of nonverbal and paralinguistic cues could reveal novel building blocks and could help building some sort of basic alphabet to identify and infer the state of the interaction. The development of suitable features and classifiers alike could increase the performance and is required for the realtime and multimodal analysis of the data.

The refinement and testing of the hierarchical architecture is an open point that needs to be addressed in the future. Datasets need to be recorded and annotated. Especially the annotation is a very time consuming procedure and could be eased by the mentioned active learning and semi-supervised learning.

As this paper provides some tools, ideas, and resources for the analysis of human behavior and the interpretation of their affective or communicational state, it is still to be positioned at the beginning of an emerging field of research, backed with several international research projects. The final solution to the very ambitious task of analyzing and interpreting the user's state in HCI is of course far from being found, however, the arguments and results of this "divide and conquer"-inspired bottom-up approach seem promising and should be investigated and evaluated in the years to come.

Acknowledgements The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl-Zeiss Foundation.

Appendix 1: Additional results

Table 3 lists a summary of the layered annotations.

Table 3 Average length (Avg. length) in seconds, standard deviation of length (StdDev. length), total duration (Tot. duration) in seconds and number of occurrences (Occurrences) of labels for user U1 and U2. All annotation layers are listed with their respective labels

Subject state	Avg. length		StdDev. length		Tot. duration		Occurrences	
	U1	U2	U1	U2	U1	U2	U1	U2
Interested	13.1	13.5	11.0	11.5	4012.6	3599.2	306	266
Uninterested	11.4	11.9	4.5	8.1	91.2	261.7	8	22
Surprised reacting	6.6	4.7	7.6	2.4	159.1	32.9	24	7
Embarrassment	9.4	9.4	7.5	7.0	366.4	103.2	40	11
Impatient	7.6	5.6	6.5	4.5	175.2	90.2	23	16
Stressed	5.8	3.5	2.7	2.3	69.5	7.0	12	2
Neg. accepting	4.9	5.1	4.0	3.3	173.0	111.5	35	22
Pos. accepting	6.1	6.1	5.0	4.7	904.5	725.9	149	119
Disagreement	5.5	9.3	3.4	5.1	82.9	102.0	15	11
Talk style								
Commanding	4.1	4.8	2.5	4.2	252.8	124.6	61	26
Off-talk	10.3	9.9	6.7	6.0	227.7	138.9	22	14
Ironic	5.4	4.0	4.9	2.8	75.2	56.0	14	14
Explaining	8.5	5.0	7.5	3.6	1296.7	474.9	153	95
Active listening	9.1	14.2	5.7	15.6	2731.5	4071.9	299	287
Question	4.3	4.4	2.5	4.3	595.6	352.0	137	80
Thinking	4.5	4.6	3.2	2.4	90.7	68.4	20	15
Reading	9.6	n/a	5.3	n/a	105.2	n/a	11	n/a
Event								
Laughs	3.2	2.9	1.5	1.7	352.9	306.8	112	107
Silence	9.0	9.3	7.2	9.4	135.1	74.6	15	8
Exciting moments	7.1	3.7	7.5	1.4	56.5	14.8	8	4
Topic shifts	2.1	2.1	1.3	1.3	21.2	21.2	10	10
Waiting	4.8	8.0	3.5	5.1	135.7	88.4	28	11

Dom. dialog role	Avg. length	StdDev. length	Total duration	Occurrences
User U1	10.9	11.2	1460.4	134
User U2	7.7	5.0	271.2	35
System S	10.5	4.8	1988.3	189
Eq. active	18.8	15.6	1765.2	94
Eq. passive	10.7	6.6	363.0	34

References

1. Aloimonos Y, Guerra-Filho G, Ogale A (2010) The language of action: A new tool for human-centric interfaces. In: Aghajan H, Augusto J, Delgado R (eds) Human centric interfaces for ambient intelligence. Elsevier, Amsterdam, pp 95–131
2. Argyle M (1988) Bodily Communication, 2nd edn. Methuen, London
3. Bachorowski J-A, Smoski MJ, Owren MJ (2001) The acoustic features of human laughter. J Acoust Soc Am 110(3):1581–1597
4. Bänziger T, Scherer KR (2007) Using actor portrayals to systematically study multimodal emotion expression: the gemep corpus. In: Proceedings of the 2nd international conference on affective computing and intelligent interaction (ACII'07). Springer, Berlin, pp 476–487
5. Batliner A, Steidl S, Eyben F, Schuller B (2010) On laughter and speech laugh, based on observations of child-robot interaction. In: The phonetics of laughing, trends in linguistics. de Gruyter, Berlin, to appear

6. Batliner A, Steidl S, Schuller B, Seppi D, Vogt T, Wagner J, Devillers L, Vidrascu L, Aharonson V, Kessous L, Amir N (2011) Whodunnit—searching for the most important feature types signalling emotion-related user states in speech. Comput Speech Lang 25(1):4–28
7. Beck C, Ognibeni T, Neumann H (2008) Object segmentation from motion discontinuities and temporal occlusions—a biologically inspired model. PLoS ONE 3(11):e3807
8. Bousmalis K, Mehu M, Pantic M (2009) Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In: Proceedings of the 3rd international conference on affective computing and intelligent interaction (ACII'09), vol 2. IEEE Press, New York, pp 1–9
9. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
10. Brennan SE (1996) Lexical entrainment in spontaneous dialog. In: Proceedings of ISSD, pp 41–44
11. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of german emotional speech. In: Proceedings of interspeech 2005, pp 1517–1520. ISCA
12. Campbell N, Kashioka H, Ohara R (2005) No laughing matter. In: Proceedings of interspeech 2005, pp 465–468. ISCA
13. Campbell WN (2004) Listening between the lines: a study of paralinguistic information carried by tone-of-voice. In: Proceedings of international symposium on tonal aspects of languages (TAL'04), Beijing, China, pp 13–16. ISCA
14. Campbell WN (2007) On the use of non verbal speech sounds in human communication. In: Lecture notes in computer science, vol 4775. Springer, Berlin, pp 117–128
15. Campbell WN (2008) Automatic detection of participant status and topic changes in natural spoken dialogues. In: Autumn meeting of the acoustical society of Japan 2008 (ASJ'08)
16. Campbell WN (2008) Tools and resources for visualising conversational-speech interaction. In: Proceedings of the sixth international language resources and evaluation (LREC'08), Marrakech, Morocco, pp 231–234. ELRA
17. Campbell WN, Scherer S (2010) Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In: Proceedings of interspeech 2010, pp 2546–2549. ISCA
18. Caridakis G, Karpouzis K, Wallace M, Kessous L, Amir N (2010) Multimodal user's affective state analysis in naturalistic interaction. J Multimodal User Interfaces 3:49–66
19. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
20. Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Commun 40(1–2):5–32
21. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. IEEE Signal Process Mag 18(1):32–80
22. Cuperman R, Ickes W (2009) Big five predictors of behavior and perceptions in initial dyadic interactions: personality similarity helps extraverts and introverts but hurts disagreeables. J Pers Soc Psychol 97(4):667–684
23. Darwin C (1978) The expression of emotion in man and animals, 3rd edn. Harper Collins, London
24. Devillers L, Vidrascu L, Lamel L (2005) Challenges in real-life emotion annotation and machine learning based detection. Neural Netw 18:407–422
25. Dietterich TG (2000) Ensemble methods in machine learning. In: Kittler J, Roli F (eds) Multiple classifier systems, proceedings of first international workshop, MCS 2000, Cagliari, Italy, June 21–23, 2000. Lecture notes in computer science, vol 1857. Springer Berlin, pp 1–15
26. D'Mello SK, Graesser AC, Schuller B, Martin J-C (eds) (2011) Proceedings of affective computing and intelligent interaction—fourth international conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Part II. Lecture notes in computer science, vol 6975. Springer, Berlin
27. Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin J-C, Devillers L, Abrilian S, Batliner A, Amir N, Karpouzis K (2007) The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In: Proceedings of the 2nd international conference on affective computing and intelligent interaction (ACII'07). Springer, Berlin, pp 488–500
28. Douxchamps D, Campbell WN (2007) Robust real time face tracking for the analysis of human behaviour. In: Lecture notes in computer science, vol 4892. Springer, Berlin, pp 1–10
29. Edlund J, Heldner M, Hirschberg J (2009) Pause and gap length in face-to-face interaction. In: Proceedings of interspeech 2009, pp 2779–2782
30. Egger S, Schatz R, Scherer S (2010) It takes two to tango—assessing the impact of delay on conversational interactivity on perceived speech quality. In: Proceedings of interspeech 2010, pp 1321–1324. ISCA
31. Ekman P (1993) Facial expression and emotion. Am Psychol 48:384–392
32. Ekman P, Friesen W (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto
33. Engel D, Spinello L, Triebel R, Siegwart R, Bülthoff H, Curio C (2009) Medial features for superpixel segmentation. In: Proceedings of the eleventh IAPR conference on machine vision applications (MVA 2009), pp 248–252
34. Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends Cogn Sci 8(1):8–11
35. Glodek M, Bigalke L, Schels M, Schwenker F (2011) Incorporating uncertainty in a layered hmm architecture for human activity recognition. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, J-HGBU'11. ACM, New York, pp 33–34
36. Glodek M, Schwenker F, Palm G (2012) Detecting actions by integrating sequential symbolic and sub-symbolic information in human activity recognition. In: International conference in stochastic modeling techniques and data analysis, to appear
37. Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, Kächele M, Schmidt M, Neumann H, Palm G, Schwenker F (2011) Multiple classifier systems for the classification of audio-visual emotional states. In: D'Mello S, Graesser A, Schuller B, Martin J-C (eds) Affective computing and intelligent interaction. Lecture notes in computer science, vol 6975. Springer, Berlin, pp 359–368
38. Gnjatovic M, Rösner D (2008) On the role of the Nimitek corpus in developing an emotion adaptive spoken dialogue system. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, Tapias D (eds) Proceedings of the sixth international language resources and evaluation (LREC'08), Marrakech, Morocco. ELRA
39. Gobl C, Bennett E, Ni Chasaide A (2002) Expressive synthesis: how crucial is voice quality? In: IEEE workshop on speech synthesis, Sept 2002. IEEE Press, New York, pp 91–94
40. Grichkovtsova I, Lacheret A, Morel M (2007) The role of intonation and voice quality in the affective speech perception. In: Proceedings of interspeech 2007, pp 2245–2248. ISCA
41. Haykin SS (1999) Neural networks: a comprehensive foundation. Prentice Hall, New York
42. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 87(4):1738–1752
43. Hermansky H (1997) The modulation spectrum in automatic recognition of speech. In: Proceedings of IEEE workshop on automatic speech recognition and understanding (ASRU'97). IEEE Press, New York, pp 140–147

44. Jaakkola TS, Haussler D (1999) Exploiting generative models in discriminative classifiers. In: Advances in neural information processing systems, pp 487–493
45. Jaeger H (2002) Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. Technical report 159, Fraunhofer-Gesellschaft, St. Augustin, Germany
46. Jokinen K, Scherer S (2012) Embodied communicative activity in cooperative conversational interactions—studies in visual interaction management. Special Issue of Acta Polytech Hung: CogInfoCom 2011 9(1):18–40
47. Kapadia S (1998) Discriminative training of hidden Markov models. PhD thesis, University of Cambridge
48. Keltner D, Ekman P, Gonzaga GC, Beer J (2003) Facial expression of emotion. In: Handbook of affective sciences. Affective science. Oxford University Press, London, pp 415–432. Chap 22
49. Kendon A (ed) (1981) Nonverbal communication, interaction, and gesture. Selections from semiotica series, vol 41. de Gruyter Berlin
50. Kennedy L, Ellis D (2004) Laughter detection in meetings. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP 2004), meeting recognition workshop. IEEE Press, New York, pp 118–121
51. Kim J, André E (2008) Emotion recognition based on physiological changes in listening music. IEEE Trans Pattern Anal Mach Intell 30(12):2067–2083
52. Kipp M (2001) Anvil—a generic annotation tool for multimodal dialogue. In: Proceedings of the European conference on speech communication and technology (Eurospeech), Aalborg, pp 1367–1370. ISCA
53. Knox M, Mirghafori N (2007) Automatic laughter detection using neural networks. In: Proceedings of interspeech 2007, pp 2973–2976. ISCA
54. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge
55. Krauss RM, Chen Y, Chawla P (1996) Nonverbal behavior and nonverbal communication: what do conversational hand gestures tell us? In: Advances in experimental social psychology. Academic Press, San Diego, pp 389–450
56. Kuncheva LI (2001) Using measures of similarity and inclusion for multiple classifier fusion by decision templates. Fuzzy Sets Syst 122(3):401–407
57. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, New York
58. Kuncheva LI, Bezdek JC, Duin RPW (2001) Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognit 34(2):299–314
59. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207
60. Lakin JL, Chartrand TL (2003) Using nonconscious behavioral mimicry to create affiliation and rapport. J Psychol Sci 14(4):334–339
61. Laskowski K (2008) Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings. In: Proceedings of the 2nd IEEE/ISCA/ACL workshop on spoken language technology (SLT'08), pp 81–84
62. Laver J (1979) The desription of voice quality in general phonetic theory. Work Prog - Univ Edinb, Dept Linguist 12:30–52
63. Laver J (1980) The phonetic description of voice quality. Cambridge University Press, Cambridge
64. Layher G, Liebau H, Niese R, Al-Hamadi A, Michaelis B, Neumann H (2011) Robust stereoscopic head pose estimation in human-computer interaction and a unified evaluation framework. In: Maino G, Foresti GL (eds) Proceedings of 16th international conference on image analysis and processing (ICIAP'11). LNCS, vol 6978. Springer, Berlin, pp 227–236
65. Lugger M, Yang B, Wokurek W (2006) Robust estimation of voice quality parameters under real world disturbances. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP 2006), vol 1. IEEE Press, New York, pp 1097–1100
66. Mackenzie Beck J (2005) Perceptual analysis of voice quality: the place of vocal profile analysis. In: Laver J, Hardcastle W, Mackenzie Beck J (eds) A figure of speech: A Festschrift for John Laver, pp 285–322. Chap 12
67. Maganti HK, Scherer S, Palm G (2007) A novel feature for emotion recognition in voice based applications. In: Proceedings of the 2nd international conference on affective computing and intelligent interaction (ACII'07). Springer, Berlin, pp 710–711
68. Mishra AK, Aloimonos Y (2009) Active segmentation. Int J Humanoid Robot 6(3):361–386
69. Monzo C, Alias F, Iriondo I, Gonzalvo X, Planet S (2007) Discriminating expressive speech styles by voice quality parameterization. In: Proceedings of 16th international congress of phonetic sciences (ICPhS'07), pp 2081–2084
70. Murphy-Chutorian MM, Trivedi E (2009) Head pose estimation in computer vision: a survey. IEEE Trans Pattern Anal Mach Intell 31(4):607–626
71. Mutch J, Lowe DG (2008) Object class recognition and localization using sparse features with limited receptive fields. Int J Comput Vis 80(1):45–57
72. Oertel C, De Looze C, Scherer S, Windmann A, Wagner P, Campbell N (2011) Towards the automatic detection of involvement in conversation. In: Esposito A, Vinciarelli A, Vicsi K, Pelachaud C, Nijholt A (eds) Analysis of verbal and nonverbal communication and enactment. The processing issues. Lecture notes in computer science, vol 6800. Springer, Berlin, pp 163–170. doi:10.1007/978-3-642-25775-9-16
73. Oertel C, Scherer S, Wagner P, Campbell N (2011) On the use of multimodal cues for the prediction of involvement in spontaneous conversation. In: Proceedings of interspeech 2011, pp 1541–1544. ISCA
74. Ogale A, Karapurkar A, Aloimonos Y (2007) View-invariant modeling and recognition of human actions using grammars. In: Dynamical vision, pp 115–126
75. Oliver N, Garg A, Horvitz E (2004) Layered representations for learning and inferring office activity from multiple sensory channels. Comput Vis Image Underst 96(2):163–180
76. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotion. Cambridge University Press, Cambridge
77. Oudeyer P-Y (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum-Comput Interact 59(1–2):157–183
78. Pantic M, Caridakis G, Andre E, Kim J, Karpouzis K, Kollias S (2011) Multimodal emotion recognition from low-level cues. In: Emotion-oriented systems, cognitive technologies. Springer, Berlin, pp 115–132
79. Pentland A (2007) Social signal processing [exploratory DSP]. IEEE Signal Process Mag 24(4):108–111
80. Pentland A (2008) Honest signals—how they shape our world. MIT Press, Cambridge
81. Petridis S, Pantic M (2009) Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: Proceedings of IEEE international conference on multimedia and expo (ICME'09). IEEE Press, New York, pp 1444–1447
82. Pickering MJ, Garrod S (2006) Alignment as the basis for successful communication. Res Lang Comput 4(2–3):203–228
83. Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifiers, pp 61–74
84. Provine R, Yong L (1991) Laughter: A stereotyped human vocalization. Ethology 89(2):115–124

85. Rabiner LR, Schafer RW (1978) Digital processing of speech signals. Prentice Hall signal processing series. Prentice Hall, New York
86. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
87. Reuderink B, Poel M, Truong K, Poppe R, Pantic M (2008) Decision-level fusion for audio-visual laughter detection. In: Popescu-Belis A, Stiefelhagen R (eds) Machine learning for multimodal interaction. Lecture notes in computer science, vol 5237. Springer, Berlin, pp 137–148
88. Russell JA (2003) Core affect and the psychological construction of emotion. Psychol Rev 110(1):145–172
89. Russell JA, Barrett LF (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 76(5):805–819
90. Schels M, Scherer S, Glodek M, Kestler HA, Palm G, Schwenker F (2011) On the discovery of events in EEG data utilizing information fusion. In: Computational statistics; special issue: Proceedings of Reisensburg 2010, pp 1–14
91. Schels M, Schwenker F (2010) A multiple classifier system approach for facial expressions in image sequences utilizing GMM supervectors. In: International conference on pattern recognition (ICPR), pp 4251–4254
92. Scherer KR, Johnstone T, Klasmeyer G (2003) Handbook of affective sciences—vocal expression of emotion. In: Affective science. Oxford University Press, London, pp 433–456. Chap 23
93. Scherer S, Glodek M, Schwenker F, Campbell N, Palm G (2012) Spotting laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. ACM Trans Interact Intell Syst: Special Issue on Affective Interact Nat Environ 2(1):4:1–4:31
94. Scherer S, Kane J, Gobl C, Schwenker F (2012) Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Comput Speech Lang. Under review
95. Scherer S, Schels M, Palm G (2011) How low level observations can help to reveal the user's state in hci. In: D'Mello S, Graesser A, Schuller B, Martin J-C (eds) Proceedings of the 4th international conference on affective computing and intelligent interaction (ACII'11), vol 2. Springer, Berlin, pp 81–90
96. Scherer S, Schwenker F, Campbell WN, Palm G (2009) Multimodal laughter detection in natural discourses. In: Ritter H, Sagerer G, Dillmann R, Buss M (eds) Proceedings of 3rd international workshop on human-centered robotic systems (HCRS'09). Cognitive systems monographs. Springer, Berlin, pp 111–121
97. Scherer S, Schwenker F, Palm G (2007) Classifier fusion for emotion recognition from speech. In: 3rd IET international conference on intelligent environments 2007 (IE'07), Sept 2007. IEEE Press, New York, pp 152–155
98. Scherer S, Strauss P-M (2008) A flexible wizard of oz environment for rapid prototyping. In: Proceedings of the sixth international language resources and evaluation (LREC'08), Marrakech, Morocco, pp 958–961. ELRA
99. Scherer S, Trentin E, Schwenker F, Palm G (2009) Approaching emotion in human computer interaction. In: International workshop on spoken dialogue systems (IWSDS'09), pp 156–168
100. Schmidt M, Schels M, Schwenker F (2010) A hidden Markov model based approach for facial expression recognition in image sequences. In: Artificial neural networks in pattern recognition (ANNPR). LNAI, vol 5998. Springer, Berlin, pp 149–160
101. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. Neural Comput 12(5):1207–1245
102. Schuller B, Valsta M, Eyben F, McKeown G, Cowie R, Pantic M (2011) In: The first international audio/visual emotion challenge and workshop (AVEC 2011). LNCS
103. Schwenker F, Dietrich C, Thiel C, Palm G (2006) Learning decision fusion mappings for pattern recognition. ICGST Int J Artif Intell Mach Learn (AIML) 6:17–21
104. Schwenker F, Sachs A, Palm G, Kestler H (2006) Orientation histograms for face recognition. In: Artificial neural networks in pattern recognition, pp 253–259
105. Schwenker F, Scherer S, Magdi Y, Palm G (2009) The GMM-SVM supervector approach for the recognition of the emotional status from speech. In: Alippi C (ed) 19th international conference on artificial neural networks 2009, Part I. LNCS, vol 5768. Springer, Berlin, pp 894–903
106. Schwenker F, Scherer S, Schmidt M, Schels M, Glodek M (2010) Multiple classifier systems for the recognition of human emotions. In: El Gayar N, Kittler J, Roli F (eds) 9th international workshop on multiple classifier systems (MCS 2010). Springer, Berlin, pp 315–324
107. Serre T, Poggio T (2010) A neuromorphic approach to computer vision. Commun ACM 53:54–61
108. Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: CVPR, pp 994–1000
109. Settles B (2009) Active learning literature survey. Computer sciences technical report 1648, University of Wisconsin-Madison
110. Shepard CA, Giles H, Le Poired BA (2001) Communication accommodation theory. Wiley, New York
111. Strauss P-M, Hoffmann H, Minker W, Neumann H, Palm G, Scherer S, Traue HC, Weidenbacher U (2008) The PIT corpus of german multi-party dialogues. In: Proceedings of the sixth international language resources and evaluation (LREC'08), Marrakech, Morocco, pp 2442–2445. ELRA
112. Strauss P-M, Hoffmann H, Neumann H, Minker W, Palm G, Scherer S, Schwenker F, Traue HC, Weidenbacher U (2006) Wizard-of-oz data collection for perception and interaction in multi-user environments. In: Proceedings of the fifth international language resources and evaluation (LREC'06), pp 2014–2017. ELRA
113. Strauss P-M, Hoffmann H, Scherer S (2007) Evaluation and user acceptance of a dialogue system using wizard-of-oz recordings. In: 3rd IET international conference on intelligent environments 2007 (IE'07). IEEE Press, New York, pp 521–524
114. Strauss P-M, Minker W (2010) Proactive spoken dialogue interaction in multi-party environments. Springer, Berlin
115. Strauss P-M, Scherer S, Layher G, Hoffmann H (2010) Evaluation of the PIT corpus or what a difference a face makes? In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the seventh international conference on language resources and evaluation (LREC'10), Valletta, Malta, pp 3470–3474. ELRA
116. Sutton C, McCallum A (2007) An introduction to conditional random fields for relational learning. In: Introduction to statistical relational learning, p 93
117. Suzuki N, Katagiri Y (2007) Prosodic alignment in human–computer interaction. Connect Sci 19(2):131–141
118. Tax DMJ, van Breukelen M, Robert W, Duin P, Kittler J (2000) Combining multiple classifiers by averaging or by multiplying. Pattern Recognit 33(9):1475–1485
119. Thiel C, Scherer S, Schwenker F (2007) Fuzzy-input fuzzy-output one-against-all support vector machines. In: 11th international conference on knowledge-based and intelligent information and engineering systems (KES'07). Lecture notes in artificial intelligence, vol 3. Springer, Berlin, pp 156–165
120. Truong KP, Van Leeuwen DA (2005) Automatic detection of laughter. In: Proceedings of interspeech 2005, pp 485–488. ISCA
121. Ullman S (1984) Visual routines. Cognition 18(1–3):97–159
122. Vaughan B (2011) Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement. In: Proceedings of interspeech 2011, pp 1865–1868. ISCA

123. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vis Comput J 27(12):1743–1759

124. Vinciarelli A, Pantic M, Bourlard H, Pentland A (2008) Social signals, their function, and automatic analysis: a survey. In: Proceedings of the 10th international conference on multimodal interfaces, ICMI'08. ACM, New York, pp 61–68

125. Walter S, Scherer S, Schels M, Glodek M, Hrabal D, Schmidt M, Böck R, Limbrecht K, Traue HC, Schwenker F (2011) Multimodal emotion classification in naturalistic user behavior. In: Jacko JA (ed) Proceedings of 14th international conference on human-computer interaction (HCII'11), vol 3. Springer, Berlin, pp 603–611

126. Watzlawick P, Beavin JH, Jackson DD (2011) Menschliche Kommunikation, Formen Stoerungen Paradoxien, 12 edn. Verlag Hans Huber, Berlin

127. Weidenbacher U, Neumann H (2009) Extraction of surface-related features in a recurrent model of v1-v2 interactions. PLoS ONE 4(6):e5909

128. Wendt B, Scheich H (2002) The "Magdeburger Prosodie Korpus"—a spoken language corpus for fMRI-studies. In: Proceedings of speech prosody 2002, pp 699–701. ISCA

129. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259

130. Wrede B, Shriberg E (2003) Spotting "hot spots" in meetings: human judgments and prosodic cues. In: Eurospeech 2003, pp 2805–2808

131. Yanushevskaya I, Gobl C, Ní Chasaide A (2005) Voice quality and f_0 cues for affect expression. In: Proceedings of interspeech, pp 1849–1852. ISCA

132. Yanushevskaya I, Gobl C, Ní Chasaide A (2008) Voice quality and loudness in affect perception. In: Proceedings of speech prosody 2008, Campinas, Brazil, pp 29–32. ISCA