Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations?

Katrin Lohan, Anna-Lisa Vollmer, Jannik Fritsch, Katharina Rohlfing, Britta Wrede

2009

Preprint:

This is an accepted article published in Int. Workshop on Social Signals Processing. The final authenticated version is available online at: https://doi.org/[DOI not available]

Proceedings

VOLUME II

Workshop on Affective Brain-Computer Interfaces
&
IEEE International Workshop on Social Signal Processing

held in conjunction with:

International Conference on
ACII 2009: Affective Computing & Intelligent Interaction
September 2009, Amsterdam, The Netherlands

IEEE Catalog Number: CFP0964H-USB
ISBN: 978-1-4244-4799-2
Library of Congress: 2009905375

Affective Brain-Computer Interfaces: Preface

This volume contains the abstracts of ABCI 2009, Affective Brain Computer Interfaces, a workshop that was organized in conjunction with ACII 2009, the International Conference on Affective Computation and Intelligent Interaction, held in Amsterdam, The Netherlands, September 2009. The workshop took place on September 9, one day before the main conference in the Keizerzaal at De Rode Hoed, Amsterdam. The workshop explored the advantages and limitations of using neurophysiological signals as a modality for the automatic recognition of affective and cognitive states, and the possibilities of using this information about the user state in innovative and adaptive applications.

Recent research in brain-computer interfaces (BCI) has shown that brain activity can be used as an active/voluntary, or passive/involuntary control modality in man-machine interaction. While active BCI paradigms have received a lot of attention in recent years, research on passive approaches to BCI still desperately needs concerted activity. More than once it has been shown that brain activations can carry information about the affective and cognitive state of a subject, and that the interaction between humans and machines can be aided by the recognition of those user states.

To achieve robust passive BCIs, efforts from applied and basic sciences have to be combined. On the one hand, applied fields such as affective computing aim at the development of applications that adapt to changes in the user states and thereby enrich the interaction, leading to a more natural and effective usability. On the other hand, basic research in neuroscience advances our understanding of the neural processes associated with emotions. Furthermore, similar advancements are being made for more cognitive mental states, for example, attention, fatigue, and work load, which strongly interact with affective states. The topics we have explored in this particular workshop are:

* emotion elicitation and data collection for affective BCI
* detection of affect and mental state via BCI and other modalities
* adaptive interfaces and affective BCI

In this workshop researchers from the communities of brain computer interfacing, affective computing, neuro-ergonomics, affective and cognitive neuroscience have been asked to present state-of-the-art progress and visions on the various overlaps between those disciplines. In addition to the paper presentations there were demonstrations by the company g.tec (Guger Technologies, Graz) and by the Frauenhofer Institute FIRST (Berlin).

The proceedings of the workshop appear as part of a volume of the ACII proceedings published by IEEE Digital Library. We are grateful to the organizers of ACII for accepting our workshop proposal. Program Chairs for ABCI2009 were Brendan Allison (TU Graz, Austria), Stephen Dunne (Starlab, Barcelona, Spain), and Dirk Heylen and Anton Nijholt, both from the University of Twente, Enschede, The Netherlands. Local chairman was Christian Muehl, also from the University of Twente. In the review process we were helped by the following members of the program committee: Anne-Marie Brouwer (TNO Human Factors, Soesterberg, The Netherlands), Peter Desain (Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands), Grandjean Didier (Swiss Center for Affective Sciences, University Geneva, Switzerland), Stephen Fairclough (School of Psychology, John Moores University Liverpool, United Kingdom), Jonghwa Kim (Institut für Informatik, Universität Augsburg, Germany), Gary Garcia Molina (Philips Research Europe, Eindhoven, The Netherlands), Femke Nijboer (Fatronik - Tecnalia, Donostia, Spain), Ioannis Patras (Department of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom), Gert Pfurtscheller (Institute for Knowledge Discovery, Technische Universität Graz, Austria), Thierry Pun (Department of Computer Science, University of Geneva, Switzerland), Egon van den Broek (Faculty of Behavioral Sciences, University of Twente, The Netherlands), and Thorsten Oliver Zander (Department Human-Machine Systems, Technische Universität Berlin, Germany).

Christian Muehl
Dirk Heylen
Anton Nijholt

Social Signal Processing Workshop: Foreword

Social interactions are among the hottest topics in the computing community. Less than a decade after the first fragmented and isolated efforts, the number of researchers active in automatic analysis, understanding and synthesis of social behavior is constantly growing and a new, vibrant research community is forming at the border between human sciences (sociology, psychology, anthropology, etc.) and technology (computer vision, speech analysis and synthesis, etc.).

Social Signal Processing is the new, emerging domain at the edge of this pioneering effort. As it establishes and formalizes for the first time a viable interface between human sciences and technology, SSP offers an ideal framework for the development of truly multidisciplinary approaches aimed at making machines socially intelligent.

The IEEE International Workshop on Social Signal Processing aims at gathering for the first time researchers approaching the problem of social intelligence in machines from all possible perspectives, namely investigation of laws and principles governing social interactions, automatic understanding of social phenomena in human-human and human-machine interactions, and synthesis of social behavior via different forms of embodiment. The goal is not only to foster cross-pollination between the above fields, but also to establish an extensive SSP community sharing common research goals and methodologies.

We take this opportunity to thank all the people that have helped to make this Workshop possible, the General Chairs of ACII 2009, the key-not speakers, the members of the Program Committee, and the reviewers. Furthermore, we acknowledge the European Network of Excellence SSPNet (www.sspnet.eu) that has supported the key-note speakers as well as the infrastructure for video recording and diffusion of all presentations.

The general chairs

Maja Pantic
Alessandro Vinciarelli

Contents

Affective Brain-Computer Interfaces: Psychophysiological Markers of Emotion in Healthy Persons and in Persons with Amyotrophic Lateral Sclerosis

Femke Nijboer Stefan P. Carmien Enrique Leon
Fabrice O. Morin Randal A. Koene Ulrich Hoffmann

Health and Quality of Life Unit, Fatronik - Tecnalia
Paseo Mikeletegi 7, 20009, Donostia - San Sebastián, Spain
fnijboer@fatronik.com

Abstract

Affective Brain-Computer Interfaces (BCI) are systems that measure signals from the peripheral and central nervous system, extract features related to affective states of the user, and use these features to adapt human-computer interaction (HCI). Affective BCIs provide new perspectives on the applicability of BCIs. Affective BCIs may serve as assessment tools and adaptive systems for HCI for the general population and may prove to be especially interesting for people with severe motor impairment. In this context, affective BCIs will enable simultaneous expression of affect and content, thus providing more quality of life for the patient and the caregiver. In the present paper, we will present psychophysiological markers for affective BCIs, and discuss their usability in the day to day life of patients with amyotrophic lateral sclerosis (ALS).

1. Toward Affective Brain-Computer Interfacing

Brain-Computer Interfaces (BCI) are systems that *measure brain signals* (*e.g.* with electroencephalogram, EEG; near-infra red spectography, NIRS; electrocorticogram, ECoG), extract *certain features* from those signals and *translate* these features into *output signals*, which are fed back (this procedure is referred to as neurofeedback[1]) to the user and/or serve as commands to control computers or machines.

BCIs and neurofeedback were first developed for treatment of medical disorders. There is substantial support for the beneficial effect of neurofeedback as a therapy for neurological disorders like epilepsy [40, 83, 84] and Attention Deficit Hyperactivity Disorder (ADHD) [2, 22, 35, 49, 78]. There is some evidence that neurofeedback is beneficial for the treatment of stroke [3, 9, 69, 85]. Furthermore, it has been suggested that neurofeedback might provide therapy for migraine [41], tinnitus [10] and personality disorders [79]. However, most neurofeedback studies (but not all [2]) tested small sample sizes and lacked a control group in which participants are given sham feedback to control for placebo effects. Thus, validation studies are needed to verify these results.

BCI research also aims to compensate for loss of motor function in people with, for example, stroke, spinal cord injury, head trauma or with neurodegenerative diseases like ALS [44]. One goal is to enable brain activity to control a robotic arm, a neuroprosthesis, or with functional electrical stimulation (FES) to control a paralyzed arm. Research focuses on invasive recording with monkeys and severely paralyzed humans [26, 36, 55, 80], and on non-invasive recording with healthy persons and those with spinal cord injury [61]. In addition, severely paralyzed patients and locked-in patients can use non-invasive BCI applications for environment control [1, 37, 64, 77] or communication programs [5,43,56,57,75,76]. Patients who are completely locked-in (lacking even the voluntary control over eye movements and of the sphincter) do not appear to be able to use a BCI [42]. Possible reasons for this go beyond the scope of this paper, but the interested reader is referred to [4,42,44].

Recently, a new perspective on BCI has emerged which suggests that not only voluntary self-regulated signals can be used as input but also that signals might tell us something about the state of the BCI user (*e.g.* the emotional and cognitive state) [18, 58, 59]. It is assumed

[1] Neurofeedback means the voluntary self-regulation of signals from the central nervous system, whereas biofeedback refers to the voluntary self-regulation of signals from the peripheral nervous system (*e.g.* electromyogramm, EMG; heart rate, HR; galvanic skin response, GSR).

978-1-4244-4799-2/09/$25.00

that relevant features from these involuntary signals (also referred to as passive signals) can be extracted and used to adapt the behavior of the HCI. Nijholt and Tan suggest that having access to the user's state is valuable to HCI and that it presents at least three distinct areas of research: 1) voluntary control over computers through brain activity, 2) evaluating interfaces and systems and 3) building adaptive user interfaces [59]. Of particular interest to HCI researchers are the user's cognitive state (*e.g*. workload of user, focus of attention) and the user's affective state (*e.g*. frustration, joy, boredom) [18]. Passive BCI could be used for healthy users and thus ease the entrance of BCIs into the market.

This notion about the passive measurement of a user's state has led to new BCI definitions [86]. First, an *active BCI* is a system that measures brain activity, extracts relevant features and translates these features into device commands or provides feedback to a user. The brain activity of the user is actively, in other words intentionally, altered. For example, the user is actively imagining opening and closing his right hand with the *intent* to alter his sensorimotor rhythm. Second, a user can be actively focusing on a certain stimulus (for example the letter "B") that he *intends* to select from a stream of stimuli (for example the whole alphabet). The desired stimulus may elicit a brain potential that can be classified by a BCI. Because the brain activity is triggered by an *exogenous* event this approach may be referred to as *reactive* BCIs. Third, a *passive BCI* (pBCI) is a system that measures ongoing, non-intentionally altered, activity from the peripheral and central nervous system, extracts relevant features and uses these features to monitor and adapt human-computer interaction. Zander and colleagues state that pBCIs are based on reactive states of the user automatically induced while interacting with a surrounding system [86].

For a schematic overview see figure 1. In this paper we aim to define affective BCI and hypothesize how to implement *affective BCIs* in healthy persons and persons with motor impairments. In our opinion, the detection of affective states begins with the discriminability of emotions, which are the smallest and most objective measurable units of affect (see section 2).

In the following paragraphs we explain the difference between emotions, feelings and moods (section 2) and introduce the field of psychophysiology in relation to emotion (section 3) and emotion theory (section 4). Furthermore, we hypothesize which psychophysiological signals might provide sensitive, reliable, and valid markers for emotion in healthy persons (section 5). Also, we explore several user scenarios in which affective BCI might be valuable for persons with motor impairment (section 6). Finally, we argue that the markers of emotion in healthy persons might be different from the markers found in persons with ALS, who are often considered as potential BCI users (section 7).

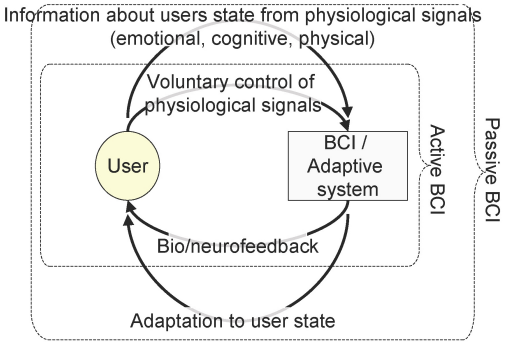


Figure 1. Schematic overview of an active and a passive BCI.

2. Emotions, Feelings and Moods

Some psychologists refer to emotion as a particular kind of subjective feeling [31], however this is a rather circular definition. In contrast, Damasio defines emotions as "*bioregulatory reactions aimed at the promotion, directly or indirectly, of the sort of physiological states that secure not just survival, but... [also] well-being*" [19]. Emotions are generally thought to be universal, short-lasting and elicited by an event, object or person. Feelings can be defined as the mental representation of the physiological changes that occur during an emotion or a mixture of emotions [19] and do not necessarily show (direct) observable peripheral reactions. In addition, a mood is a sustained tendency toward certain emotions (*e.g*. depression). From this point of view, for example, fear would be an emotion, restlessness a feeling and anxiety a mood. The whole range of emotions, feelings and moods may be called affect. Although Damasio has received some criticism [32], the neurobiological perspective of his definition seems to offer the best starting point for affective Brain-Computer Interfacing, which aims at classifying emotional states without verbally asking the subject about his or her subjective feeling.

There are two important issues that are worth highlighting in relation to the study of emotions in the context of BCI and affective computing. First, to advance the modelling of emotions by means of computing systems, researchers should not wield or attach to a particular theory or definition of emotion. The rationale for this is that the discussion on the meaning of emotions is an ongoing theoretical controversy that in 1981 had already yielded 92 different descrip-

tions of emotions [39]. Instead, technologists should work from a basis of the widespread view of emotions as a multi-element phenomenon that involves a) appraisal of events, b) psychophysiological changes, c) motor expressions, d) action tendencies, e) subjective experiences, and f) emotion regulation [27]. Affective BCI should focus on those elements of emotion that are easy to measure or to synthesize such as motor expressions, actions or physiological activation. Second, computer systems are still highly dependent on data acquired from a number of individuals who are subjected to certain type of emotional stimulation. Thus, the method to elicit emotions under controlled laboratory conditions is as important as the techniques employed to detect, classify or simulate affective states. Although not a single elicitation method can guarantee that a given targeted emotional state or class is experienced, some instruments have been shown to work well under certain circumstances(*e.g.* films, music, scripted interactions).

In this context, we argue that there are two main approaches to the study of emotions that seem to fit well with the aims of affective BCI. On the one hand Ekman's emotional classification or factorial approach represents a rather balanced way to endow subjective levels to a number of emotional states without getting into the controversy of whether there are two, twelve or more identifiable affects. Ekman's work has been traditionally associated with the use of facial expressions in emotion detection (for the corresponding facial expressions see figure 2). Ekman listed joy, sadness, fear, anger, surprise and disgust as the six basic emotions [23] (for the corresponding facial expressions see figure 2 below).

On the other hand, dimensions are very useful to quantify elements of emotions without the need to utilize pre-defined labels. The "bi-phasic model of emotion", which was proposed by Lang and colleagues, emerges from a motivational perspective that points to emotion as a behavioral tendency of a subject to approach or avoid/withdraw from a stimulus [6, 47, 72]. Emotions can be organized as pleasant/appetitive versus unpleasant/aversive and this disposition constitutes the first bipolar dimension of the model - valence. In addition, emotions can mobilize energy to different degrees, and therefore the activation or the intensity can vary. The model hereby constitutes a second bipolar dimension - arousal. An additional bipolar dimension - dominance-submissiveness - has been proposed to measure emotion [70]. However, valence and arousal level explain the greater portion of the variance in emotion [71]. For two reasons we prefer the bi-phasic model of emotion as opposed to approaches which describe four [27] or more dimensions [16]. First, Lang's two dimensions facilitate experimentation because they are applicable to a variety of affective phenomena

Figure 2. The facial expressions belonging to each basic emotion defined by Ekman [23]. From left to right; top: anger, joy, disgust; bottom: surprise, sadness, fear.

and second, they are also closely linked to a very popular elicitation method that employs a standardized set of photographs, the International Affective Picture System [46]. It is worth mentioning that the use of a factorial and/or dimensional approach to measuring emotions has also been suggested in the context of affective pervasive systems (see for example [50] and [53]).

Emotions elicited by stimuli can be rated within the valence-arousal space by using the Self Assessment Manikin (SAM) (see figure 3) [7]. SAM is a non-verbal graphical tool on which subjects have to rate on a nine-point scale how they feel. Valence is depicted as a smiling happy, figure transitioning into a frowning, unhappy figure. For arousal SAM ranges from a sleepy figure, with eyes closed, to an excited figure, with eyes open. Because SAM is a language-free, culture-free measurement it is suitable for various countries. However, before one can rate the emotion that was elicited by a stimulus, a mental reflection on this emotion is required. Thus, according to the strict definition of Damasio, one would have to say that the SAM measures feelings and not emotions. For affective BCI research however, the correlation of subjective feelings (as measured by the SAM) to psychophysiological signals (*e.g.* EEG, EMG) might result in a sensitive, reliable and valid model for affective BCI applications.

3. An Illustration from History

The search for reliable and objective indicators of emotional states stretches back as far as the period of 290 to 280 B.C. [74]. Antiochus of Apama, son of king Seleucus I, found himself hopelessly in love with his stepmother, a young woman by the name of Stratonic. Antiochus, an obedient and submissive son, fought with all his might

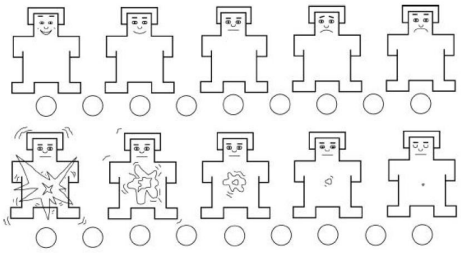


Figure 3. The self-Assessment Manikin (SAM). Top: valence; bottom: arousal

against these feelings and never spoke a word with anyone about the matter. He suffered so much from his love sickness that he became seriously ill and was brought to Eristratus, a grandson of Aristotle, who was very well educated. Plutarch wrote: "*Eristratus, the 'medical man', understood without difficulty that Antiochus was in love, but as he wanted to find out who was the object of his passion - not an easy task - he installed himself in Antiochus' chamber, living therein. Whenever a goodlooking girl or a youth appeared before them, he keenly watched Antiochus' face in order to discover signs of emotions or change of expression. He also watched his body, looking out for any movements of his limbs and body or alterations of the same, which are naturally affected when the soul is under violent states. He was thus able to establish that no change was produced in Antiochus, excepting whenever Stratonic appeared, either alone or in Seleucus company. Sappho's symptoms became then all too apparent, such as a break in the voice, blushing and downcast eyes, sudden perspiration and irregularity of the pulse. He also became subject to swoons, doubts, fears, and sudden pallor. From all these manifestations Eristratus drew the conclusion that the king's son loved nobody but her, and that he was determined rather to die than to show it*" [65].

Eristratus classified emotions based on their co-occurrence with stimuli (independent variable: beautiful women; see figure 4). He operationalized emotion with the following dependent variables: voice quality, eye movements, skin responses, and blood pressure. This may have constituted the first documented psychophysiological study. It illustrates how the classification of emotion is important for understanding how emotions change our perception, guide our behavior, and shape our memory. Emotion detection and mimicry is an important requirement for maintaining successful social relations with others. However, *whether* emotions can be distinguished based on differences in the activity of the central and autonomic nervous systems is a highly debated topic in emotion theory [13, 38, 73]).

Figure 4. Eristratus classifies the cause of the illness in Antiochus. A painting by Jacques-Louis David.

4. Theories of Emotion

William James and Carl Lange simultaneously and independently hypothesized in 1890 that contrary to common belief "*the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion*" [38]. James states for example that we do not flee because we are afraid when we see a bear, but we are afraid because we flee from the bear. Similarly, we do not cry because we feel sad after bad news, but we feel sad because we are crying. The James-Lange hypothesis, also referred to as a a peripheral theory of emotion, implies that emotions can be differentiated by somatovisceral responses. However, bodily changes are not consistently associated with specific emotions (see section 5). The hypothesis also implies that people with quadriplegia should not show any emotional responses, which is refuted by several studies [15].

In 1928 Walter Cannon presented a critical examination of the by then popular James-Lange notion on emotion [13]. He postulated his own theory that the viscera and the innervation of the muscles were not the sources for the qualities of emotion. He held that emotions are derived from subcortical centers (*e.g.* thalamus) and that peripheral activity is not necessary for emotional experience. In other words the sight of a bear can cause fear without fleeing. Support for this theory comes from studies that show that direct brain stimulation can cause emotion experience. The Cannon's theory is sometimes referred to as a centralistic theory.

Another important emotion theory was proposed by Schachter and Singer, who suggested bodily changes qualify as emotions only when coupled with judgements that attribute these changes to emotionally relevant objects or events (this process is also referred to as appraisal) [73].

When our heart beats fast in the presence of a bear, we would attribute (appraise) that bodily change to the bear and feel afraid. In contrast, when our heart beats fast in the presence of an attractive person in the same room, we would attribute that bodily change to lust or love. Thus, Schachter and Singer state that bodily changes are essential but not sufficient.

The above mentioned theories are but few among many emotion theories. We refer the interested reader to [51, 60]. The debate in emotion theory is of high relevance to the area of affective Brain-Computer Interfacing, since this area will depend on at least some degree of distinct visceral or brain patterns underlying different emotions. On the other hand the technologies and methods developed by the BCI field might contribute to new approaches for emotion classification and might lead to a more multidisciplinary field of emotion research. In the next section we will review the evidence for the discriminability of various emotions within the EEG and some peripheral measures.

5. Psychophysiological Markers of Emotions in Healthy Persons

Emotion, defined as bioregulatory reactions [19], can be studied through psychophysiological signals from the central and the peripheral nervous system, through audio-recordings of speech signals, and through video-recordings of facial expressions. There is extensive literature about emotion assessment from audio- and video-recordings. However, these two modalities have the disadvantage that they require the active participation of the user (speak, or look into the camera) and hence cannot be measured continuously and reliably.

A literature search shows that relatively few peer-reviewed papers exist about the classification of emotions based on signals from the central nervous system. This is most probably due to the fact that it is very difficult to reliably classify emotions from non-invasively acquired brain signals such as the EEG. An exception to the scarcity of literature in this area is the line of work of Davidson et al [20] in which it is extensively argued that the prefrontal cortex plays an important role in emotional processing. In particular, hemispheric differences in alpha-power over the frontal cortex are repeatedly mentioned as indicator for emotions.
From a more practical point of view Chanel and colleagues compared three approaches to classify 3 emotions in 10 participants [14]. The classification was performed using data from 1) only EEG signals, 2) only peripheral signals or 3) a combination of both types of signals. They report classification accuracies between 50% and 65% for classification based on either peripheral signals or EEG signals and a classification accuracy of about 70% for combining both modalities. The combination of signals from the peripheral and central nervous system thus seems promising. This is the reason why an affective BCI system should draw not only on EEG signals but also peripheral signals.

A vast amount of literature exists about the assessment of emotions based on psychophysiological measures from the peripheral nervous system. Examples of psychophysiological measures are electromyogram (EMG), skin conductivity (*e.g.* galvanic skin response; GSR), heart rate (HR), heart rate variability (HRV), blood pressure (BP), and respiration (RSP). One of the first ways to measure emotion is to instrument the muscles in the face that are responsible for facial expressions that are obvious reflections of emotions. There is a large body of research [12, 33, 82] attempting to tie specific muscle sets to types of emotions; electromyogram (EMG) of the facial muscles in specific and of others, both measuring general arousal [81] and specific indicators [48]. Another physical measurement that ties to emotional state is the heart rate, which is a good measure of arousal [12]. Skin conductivity reflects the outputs of the eccrine sweat glands, which reside on the palms of the hands and the soles of the feet and are particularly responsive to emotional activation, and only minimally responsible for thermoregulation [63, 81]. Also commonly used are blood pressure and respiration [11]. Less commonly used are such measures as pupil dilation [30], posture [21], cardiac output, diastolic blood pressure, eye blink rate, face temperature, finger temperature, heart rate variability, number of muscle tension peaks, oxygen saturation of the blood, and inspiratory time [12]. All of these can be used in combination to produce classifiers of affective states. Cacioppo and colleagues have provided an extensive meta-review of the literature in 1998 [12].

With a set of inputs (from some grouping of the psychophysiological data described above) and a list of classes (emotions) the next part of emotion recognition is the process of classification. Just like any classification problem the steps are signal acquisition, signal conditioning, feature extraction, training and finally producing a classification function. The raw data of the psychophysiological signals are typically taken as a value that is part of a waveform and then normalized and combined in various permutations and with various feature extraction functions [63]. The next step is to reduce the number of dimensions given to the classification algorithm (to reduce the possibility of overfitting the classifier to the training data) [62]. Algorithms used in classifying span from sequential floating

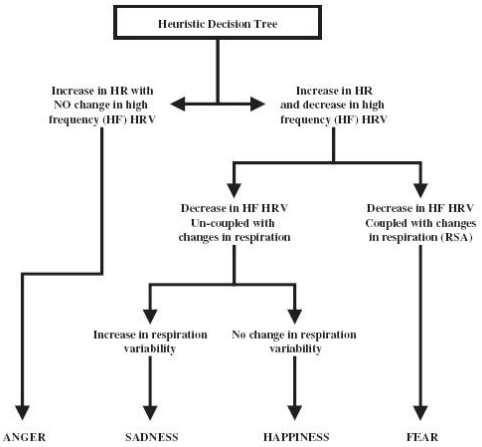


Figure 5. Heuristic Decision Tree based on heart rate and heart rate variability. Taken from [68].

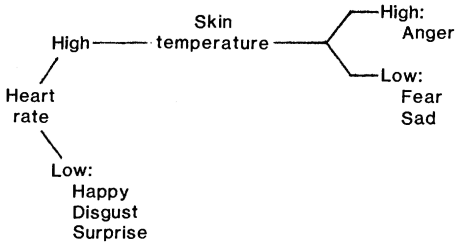


Fig. 2. Decision tree for discriminating emotions in direction facial action task.

Figure 6. Heuristic Decision Tree based on heart rate and skin temperature. Taken from [24].

forward search and Fisher projection and a permutation of both [11] to neural networks [8] and hidden markov models [7]. Several papers presented heuristic decision trees for classification, one based on heart rate and heart rate variability [68], and another on heart rate and skin temperature [24] (see figure 5 and 6 below).

How accurately a classifier can identify an emotion solely on the basis of psychophysiological data is dependent on the selected sensors, the classification process and several other parameters (which will be described below). The studies referenced in this paper obtained an accuracy range spanning 65.3 where 25 % would be chance [68] to 76.8 % [62] and 50.62 % [82] where 12.5 %would be chance, to 89.73 %, 63.76 % [33], and 63.4 % [25] where 50 % would be chance. The studies used both different lists of emotions (both in number and content) as well as psychophysiological data so this list then is of use in confirming that emotions can be automatically recognized with some degree of confidence. Similarly, determining the optimal combination of sensors and features extracted that can best classify the presence of a given basic emotion is a goal that needs to be reached through empirical approaches in which scientists from emotion psychology, affective neuroscience, brain-computer interface and neuroinformatics should closely work together.

A review of the literature also returned several concerns that are important to keep in mind in designing BCI studies with respect to psychophysiological markers. In many cases a classifier trained on a single person will not accurately classify signals from another person, therefore every subject may need to have an individually trained classifier [63]. Secondly, it has been noted that "the features extracted from the signals are highly dependent on the day the experiment was held" [62]. Thus, it may be necessary to create a new classifier (or at least regenerate the features) for each subject and each session. Thirdly, research points out that an individual's psychophysiological response to a given emotion changes as they age [12]. Fourthly, psychophysiological markers of emotion can be easily confounded by external factors (*e.g.* day light, temperature, body position, time of day), substance intake (*e.g.* nicotine, caffeine, high caloric food) and physical activities. Technological solutions to measure these changes in the environment may include light sensors, accelerometers or a thermometer. Fifthly, multimodal classification methods need to applied to these various signals and compared. Sixthly, one will want to know whether a change in psychophysiological signals reflects an emotion (phasic change) or a steady state (tonic change). For example, a low blood pressure may indicate low emotional arousal, but it may also indicate a person is asleep. An ideal affective BCI classifier would have knowledge of time and events in the environment of the user (*e.g.* someone entered the room, there is a storm outside, time since last shower).

6. Affective BCI for Persons with Amyotrophic Lateral Sclerosis

ALS is a fatal motor neuron disease of unknown etiology and cure. ALS is a neurodegenerative disorder of large motor neurons of the cerebral cortex, brain stem, and spinal cord that results in progressive paralysis and wasting of muscles [17]. ALS has an incidence of 2/100.000 and a prevalence of 6-8/100.000 [8]. Survival is limited by respiratory insufficiency. Most patients die within 3-5 years after onset of the disease [17], unless they choose life-sustaining treatment [34].

Figure 7. Schematic overview of an active and a passive BCI for people with motor impairment. Information of the user state is fed back to the caregiver with whom the user is interacting.

As the disease progresses, people become increasingly paralyzed. The first symptoms experienced by most patients include weakness in arms or legs, after which the paralysis spreads to other extremities and finally also the neck and head areas. This form of ALS is called spinal ALS. On the contrary, bulbar ALS starts with symptoms of weakness and paralysis in neck and mouth regions and then spreads to other extremities. Involuntary muscle contractions in late-stage ALS can occur during emotional experience.

An illustrative example is given from a visit from the first author (FN) to HPS, the patient who was the first to use a BCI in his daily life for communication [5]. HPS was locked-in at the time of the visit. He could raise his eye brow to say 'no' and half-close his eyes to say 'yes'. During this visit FN and HPS did not use a BCI to communicate but instead a caregiver served as an interlocutor. First, the caregiver read out loud the number of the rows in a letter matrix until HPS selected the row containing his desired letter. Then, the caregiver read out loud the letters in that row until HPS selected his desired letter. This procedure repeated itself until words and sentences were formed. HPS, being German, asked how FN, being Dutch, felt about an upcoming soccer match between the Netherlands and Germany in the following week. FN replied she was certain that the Netherlands were going to win and that "it would be a piece of cake". This remark appeared to elicit two emotions in HPS. First, he smiled involuntarily. Second, his eyes peered attentively to FN, who interpreted these expressions as an indication that HPS wanted to reply with a furious yet witty remark.

However, humor, happiness and anger, are very difficult for severely paralyzed patients to express. Even though HPS dictated his reply, he lacked the ability to modulate the tone of his voice or use his facial expression to add sarcasm. From this example a first purpose of an affective BCI can be identified: they may offer a possibility to otherwise poker-faced patients to express their affect. Figure 7 illustrates how an affective BCI might not only adapt HCI for a patient, but also provide information about the affective state of the user to a caregiver, who is interacting with the user. From our experience we know that caregivers often leave the room while patients 'write' lengthy messages with their assistive technology, only to come back when the whole sentence is written down. Sometimes messages go unnoticed or the context of the message may be forgotten by the time the message is written. Receiving nonverbal input from a patient may provide context and constitute an important incentive to continue interacting with the patient, especially when content is conveyed slowly. Also, perception of the affective state of the user may cause mimicry of these states in the caregiver, reassuring the patient that he or she is perceived and understood. We hypothesize that affective BCI will improve the quality of life and interaction of patients and caregivers, because *affect* and *content* can be *simultaneously* expressed. An application of such an affective BCI could be a monitor attached to the patient's wheelchair displaying a face expressing the emotions detected by the algorithms.

Finally, an emotion detection system could also serve as an alarm system to cue the caregiver to check on a patient. Although medical devices surrounding the patients (*e.g.* artificial respiration) measure heart rate and blood oxygen level and give an alarm when for example blood oxygen level is too low, psychological distress does not give an alarm. Thus, a paralyzed patient is rendered powerless when a frightening event happens. An affective BCI might detect from GSR and heart rate that negative emotions with strong arousal are felt by the patient and send an alarm signal to the caregiver. However, in the next section we will discuss how emotional markers might be different in patients with ALS compared to healthy controls.

7. Emotional Processing in Patients with Amyotrophic Lateral Sclerosis

There are only few studies on affect and emotional processing in ALS. Remarkably few patients (9-11 %) develop a major depressive disorder despite the severe impact the disease has on a persons life [28, 45, 66, 67]. Lulé and colleagues investigated emotional processing in ALS [52]. Twelve ALS patients and eighteen age-matched healthy controls were (neuro)psychologically assessed. Then, they rated their emotions with the SAM after viewing negative, neutral and positive pictures from the International Affective Picture System (IAPS) [46]. In a second experiment physiological responses to the same pictures

were measured. Specifically, startle response and heart rate were measures as an index of valence and galvanic skin response as an index of arousal.

Compared to controls, patients rated positive and neutral pictures as more positive and negative pictures as less negative. Also, calm and neutral stimuli were rated as more arousing, whereas most arousing pictures (especially those with erotic content) were rated as less arousing. GSR were significantly delayed compared to controls, while the amplitude of GSR tended to be higher in ALS than in healthy controls. Both ALS patients and healthy controls showed a stronger HR deceleration after unpleasant stimuli compared to after pleasant stimuli.

The altered rating of emotional stimuli was not correlated to depression scores or frontal lobe dysfunction. Lulé and colleagues therefore suggest that emotional processing is altered due to coping mechanisms. None of the ALS participants in this study was locked-in and little is known about emotional processing in patients with late-stage ALS.

Moore and Dua provided many biofeedback training sessions to a locked-in patient with ALS. The patient was progressing to the complete locked-in state (no voluntary eye movement or sphincter control) during the experiment, which lasted over a year [54]. The patient aimed to learn to say 'yes' by raising his GSR level and to say 'no' by keeping the GSR level low. After a year the accuracy of saying 'yes' and 'no' was significantly above chance level, but probably not sufficient to reliably answer questions. GSR differs between ALS patients and healthy persons [29] and it remains questionable if self-regulation of GSR might be used for communication in ALS patients and how GSR can be used for emotion detection in ALS patients.

Furthermore, patients with motor impairment might depend on life-sustaining devices, like artificial respiration or percutaneous endoscopic gastrostomy (PEG), that may affect the peripheral and central nervous system. Also, medication, like antidepressants or diabetes medication, may cause affective states in patients to be differently classified. Finally, it must be noted that an interesting line of investigation might be to study whether facial expression that is not overtly observable might be detected by EMG measurements in severely paralyzed patients patients. Few and potentially inexpensive electrodes might classify the valence of emotions in these patients.

8. Conclusion

The concept of affective and passive BCIs has lead to a new perspective on the applicability of BCIs. Affective BCIs may now serve as assessment tools for HCI and adaptive system to improve HCI with healthy people. Affective states should be measured through a synthesis of peripheral and central measures although a solution of the optimal parameters is still not present. Also, it may be discussed whether the term *brain*-computer interface is then still appropriate or if we should find a more generic term such as body-computer interface or even human-computer interface.

Affective BCIs may improve the quality of life of persons with motor impairment and of their caregivers, by allowing the BCI user to express not only content but also affect. However, the accurate detection of affect is not a simple matter, and successful approaches with patients may differ from those used in healthy persons.

References

[1] F. Aloise, F. Cincotti, F. Babiloni, M. G. Marciani, D. Morelli, S. Paolucci, G. Oriolo, A. Cherubini, F. Sciarra, F. Mangiola, A. Melpignano, F. Davide, and D. Mattia. The ASPICE project: inclusive design for the motor disabled. In *Proceedings of the Working conference on Advanced Visual Interfaces*, pages 360 – 363l, 2006.

[2] M. Arns, S. de Ridder, U. Strehl, M. Breteler, and A. Coenen. Efficacy of neurofeedback treatment in ADHD: The effects of inattention, impulsivity and hyperactivity: a meta-analysis. *Clinical EEG and Neuroscience*, 40:3, 2009.

[3] T. Bearden, J. Cassisi, and M. Pineda. Neurofeedback training for a patient with thalamic and cortical infarctions. *Applied Psychophysiology Biofeedback*, 28(3):241–253, 2003.

[4] N. Birbaumer. Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.

[5] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[6] M. Bradley. *Handbook of Psychophysiology*, chapter Emotion and motivation, pages 602–642. Cambridge University Press, 2000.

[7] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[8] B. Brooks. Clinical epidemiology of amyotrophic lateral sclerosis. *Neurological Clinics*, 14(2):399–420, 1996.

[9] E. Buch, C. Weber, L. Cohen, C. Braun, M. Dimyan, T. Ard, J. Mellinger, A. Caria, S. Soekadar, A. Fourkas, and N. Birbaumer. Think to move: A neuromagnetic brain-computer interface (BCI) system for chronic stroke. *Stroke*, 39(3):910–917, 2008.

[10] M. Busse, Y. F. Low, F. I. Corona-Strauss, W. Delb, and D. J. Strauss. Neurofeedback by neural correlates of auditory selective attention as possible application for tinnitus therapies. *Conference Proceedings IEEE Engineering in Medical and Biological Society*, 2008:5136–5139, 2008.

[11] J. Cacioppo, G. Berntson, J. Larsen, K. Poehlmann, and T. Ito. *The Handbook of Emotion*, chapter The psychophysiology of emotion, pages 173–191. The guilford press, 2000.

[12] J. Cacioppo, G. Berntston, D. Klein, and K. Poehlmann. *Annual Review of Gerontology and Geriatrics*, chapter The psychophysiology of emotion across the lifespan, pages 27–65. Springer Publishing Company, 1998.

[13] W. B. Cannon. The James-Lange theory of emotions: a critical examination and an alternative theory. *American Journal of Psychology*, 100(3-4):567–586, 1927.

[14] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67:607 – 627, 2009.

[15] K. Chwalisz, E. Diener, and D. Gallagher. Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured. *Journal of Personality and Social Psychology*, 54(5):820–828, May 1988.

[16] T. Cochrane. 8 dimensions for emotions. *Social Science Information. Special issue on The language of emotion: conceptual and cultural issues*, 48(3), 2009.

[17] M. Cudkowicz, M., and J. Shefner. Measures and markers in amyotrophic lateral sclerosis. *NeuroRx*, 1(2):273–83, 2004.

[18] E. Cutrell and D. Tan. BCI for passive input in HCI. In *Proceedings of CHI 2008*. Microsoft Research, ACM, 2008.

[19] A. Damasio. *Feelings and emotions: The Amsterdam symposium*, chapter Emotions and feelings - A neurobiological perspective, pages 49–57. Cambridge University Press, 2004.

[20] R. Davidson. Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology*, 40:655 – 665, 2003.

[21] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*. AMC Press, 2005.

[22] R. Drechsler, M. Straub, M. Doehnert, H. Heinrich, H.-C. Steinhausen, and D. Brandeis. Controlled evaluation of a neurofeedback training of slow cortical potentials in children with Attention Deficit/Hyperactivity Disorder (ADHD). *Behavioral and Brain Functions*, 3:35, 2007.

[23] P. Ekman, W. Friesen, and P. Ellsworth. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Pergamon Press, 1972.

[24] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983.

[25] R. Fernandez and R. Picard. Signal processing for recognition of human frustration. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, WA, 1998.

[26] E. E. Fetz and D. V. Finocchio. Operant conditioning of specific patterns of neural and muscular activity. *Science*, 174(7):431–435, 1971.

[27] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, Dec 2007.

[28] L. Ganzini, W. S. Johnston, and W. F. Hoffman. Correlates of suffering in amyotrophic lateral sclerosis. *Neurology*, 52(7):1434–40, 1999.

[29] N. Ghanayim. *Psychophysiologische Untersuchungen bei Patienten mit amyotropher Lateralsklerose (ALS) zur Differenzierung von Emotionskonzepten*. Master thesis, University of Tübingen, 2000.

[30] E. Granholm and S. R. Steinhauer. Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, 52(1):1–6, 2004.

[31] P. Gray. *Psychology*. Worth publishers, 1994.

[32] P. Griffiths. *What emotions really are: the problem of psychological categories*. University of Chicago Press, illustrated edition, 1997.

[33] A. Haag, S. Goronzy, P. Schaich, and J. Williamsh. Using bio-sensors: First steps towards an automatic system. In *Workshop on Affective Dialogue Systems*, 2004.

[34] H. Hayashi and E. A. Oppenheimer. ALS patients on TPPV: Totally locked-in state, neurologic findings and ethical implications. *Neurology*, 61(1):135–7, 2003.

[35] H. Heinrich, H. Gevensleben, F. Freisleder, G. Moll, and A. Rothenberger. Training of slow cortical potentials in Attention-Deficit/Hyperactivity Disorder: Evidence for positive behavioral and neurophysiological effects. *Biological Psychiatry*, 55(7):772–775, 2004.

[36] L. Hochberg, M. Serruya, G. Friehs, J. Mukand, M. Saleh, A. Caplan, A. Branner, D. Chen, R. Penn, and J. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

[37] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens. An efficient P300-based brain-computer interface for disabled subjects. *Journal of Neuroscience Methods*, 167(1):115–125, 2008.

[38] W. James. What is an emotion? *Mind*, 9:188–205, 1890.

[39] P. Kleinginna and A. M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, 1981.

[40] B. Kotchoubey, U. Strehl, S. Holzapfel, V. Blankenhorn, W. Fröscher, and N. Birbaumer. Negative potential shifts and the prediction of the outcome of neurofeedback therapy in epilepsy. *Clinical Neurophysiology*, 110(4):683–686, 1999.

[41] P. Kropp, M. Siniatchkin, and W.-D. Gerber. On the pathophysiology of migraine–links for "empirically based treatment" with neurofeedback. *Applied Psychophysiology Biofeedback*, 27(3):203–213, 2002.

[42] A. Kübler and N. Birbaumer. Brain-computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients? *Clinical Neurophysiology*, 119(11):2658–2666, 2008.

[43] A. Kübler, N. Neumann, J. Kaiser, B. Kotchoubey, T. Hinterberger, and N. P. Birbaumer. Brain-computer communication: Self-regulation of slow cortical potentials for verbal communication. *Archives of Physical Medicine and Rehabilitation*, 82(11):1533–1539, 2001.

[44] A. Kübler, F. Nijboer, and N. Birbaumer. Brain-Computer Interfaces for Communication and Motor Control - Perspectives on Clinical Applications. In G. Dornhege, J. Millan,

T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing*, pages 373–392. The MIT Press, 2007.

[45] A. Kurt, F. Nijboer, T. Matuz, and A. Kübler. Depression and anxiety in individuals with amyotrophic lateral sclerosis: Epidemiology and management. *CNS Drugs*, 21(4):279–291, 2007.

[46] P. Lang, A. Öhman, and D. Vaitl. The international affective picture system [photographic slides]. Technological report, The Center for Research in Psychophysiology, University of Florida, Gainsville, FL, 1988.

[47] P. J. Lang. The emotion probe. studies of motivation and attention. *American Psychologist*, 50(5):372–385, May 1995.

[48] J. Larsen, G. Berntson, H. Poehlmann, T. Ito, and J. Cacioppo. *The handbook of emotions*, chapter The psychophysiology of emotion, pages 180–195. Cambridge University Press, 2008.

[49] U. Leins, G. Goth, T. Hinterberger, C. Klinger, N. Rumpf, and U. Strehl. Neurofeedback for children with ADHD: a comparison of SCP and Theta/Beta protocols. *Applied Psychophysiology Biofeedback*, 32(2):73–88, 2007.

[50] E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda. A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering Applications of Artificial Intelligence, The International Journal of Intelligent Real-Time Automation*, 20(3):337–345, 2007.

[51] M. Lewis and J. Haviland-Jones, editors. *The handbook of emotion*. The guilford press, second edition edition, 2000.

[52] D. Lul, A. Kurt, R. Jürgens, J. Kassubek, V. Diekmann, E. Kraft, N. Neumann, A. Ludolph, N. Birbaumer, and S. Anders. Emotional responding in amyotrophic lateral sclerosis. *Journal of Neurology*, 252(12):1517–1524, 2005.

[53] I. Montalban, A. Garzo, and E. Leon. Emotion-aware intelligent environments: A user perspective. In *Proceedings of the International Conference on Intelligent Environments*, 2009.

[54] M. Moore and U. Dua. A galvanic skin response interface for people with severe motor disabilities. *ACM SIGACCESS Accessibility and Computing*, (77-78):48–54, 2004.

[55] C. Moritz, S. Perlmutter, and E. Fetz. Direct control of paralysed muscles by cortical neurons. *Nature*, 456(7222):639–642, 2008.

[56] C. Neuper, G. R. Mller, A. Kbler, N. Birbaumer, and G. Pfurtscheller. Clinical application of an EEG-based brain-computer interface: A case study in a patient with severe motor impairment. *Clinical Neurophysiology*, 114(3):399–409, 2003.

[57] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D. J. McFarland, N. Birbaumer, and A. Kübler. An auditory brain-computer interface (BCI). *Journal of Neuroscience Methods*, 167(1):43–50, 2008.

[58] A. Nijholt, B. Allison, M. Jackson, D. Tan, J. Milln, and B. Graimann. Brain-computer interfaces for HCI and games. In *28th Annual CHI Conference on Human Factors in Computing Systems*, pages 3925–3928, 2008.

[59] A. Nijholt and D. Tan. Playing with your brain: Brain-computer interfaces and games. Organizing paper for Brainplay Workshop at Advances in Computer Entertainment, 2007.

[60] W. Parrott. *Emotions in social psychology*. Psychology Press, 2000.

[61] G. Pfurtscheller, C. Guger, G. Müller, G. Krausz, and C. Neuper. Brain oscillations control hand orthosis in a tetraplegic. *Neuroscience Letters*, 292(3):211–214, 2000.

[62] R. Picard and E. Vyzas. Affective pattern classification. In *AAAI Fall symposium series. Emotional and intelligent: The tangled knot of cognition*, 1998.

[63] R. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1175–1191, 2001.

[64] F. Piccione, F. Giorgi, P. Tonin, K. Priftis, S. Giove, S. Silvoni, G. Palmas, and F. Beverina. P300-based brain computer interface: reliability and performance in healthy and paralysed participants. *Clinical Neurophysiology*, 117(3):531–537, 2006.

[65] Plutarch. *Las vidas paraleles*. Libreria de Hernando y Cia., 1901.

[66] J. Rabkin, S. Albert, M. D. Bene, I. O'Sullivan, T. Tider, L. Rowland, and H. Mitsumoto. Prevalence of depressive disorders and change over time in late-stage ALS. *Neurology*, 65(1):62–7, 2005.

[67] J. Rabkin, S. Albert, T. Tider, M. D. Bene, I. O'Sullivan, L. Rowland, and H. Mitsumoto. Predictors and course of elective long-term mechanical ventilation: A prospective study of ALS patients. *Amyotrophic Lateral Sclerosis*, 7(2):86–95, 2006.

[68] P. Rainville, A. Bechara, N. Naqvi, and A. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61(1):5–18, 2006.

[69] G. R. Rozelle and T. H. Budzynski. Neurotherapy for stroke rehabilitation: A single case study. *Biofeedback and Self-Regulation*, 20(3):211–228, 1995.

[70] J. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[71] J. A. Russell and L. F. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819, 1999.

[72] J. Sanchez-Navarro, J. Martínez-Selva, G. Torrente, and F. Román. Psychophysiological, behavioral, and cognitive indices of the emotional response. *The Spanish Journal of Psychology*, 11(1):16–25, 2008.

[73] S. Schachter and J. E. Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69:379–399, 1962.

[74] C. A. Seguin. Erasistratus, antiochus, and psychosomatic medicine. *Psychosomatic Medicine*, 10(6):355, 1948.

[75] E. Sellers and E. Donchin. A P300-based brain-computer interface: initial tests by ALS patients. *Clinical Neurophysiology*, 117(3):538–548, 2006.

[76] E. Sellers, A. Kübler, and E. Donchin. Brain-computer interface research at the University of South Florida Cognitive Psychophysiology Laboratory: the P300 speller. *IEEE*

Transactions on Neural Systems and Rehabilitation Engineering, 14(2):221–224, 2006.

[77] S. Silvoni, C. Volpato, M. Cavinato, M. Marchetti, K. Priftis, A. Merico, P. Tonin, K. Koutsikos, F. Beverina, and F. Piccione. P300-based brain-computer interface communication: evaluation and follow-up in amyotrophic lateral sclerosis. *Frontiers in Neuroprosthetics*, 2009 (in press).

[78] U. Strehl, U. Leins, G. Goth, C. Klinger, T. Hinterberger, and N. Birbaumer. Self-regulation of slow cortical potentials: A new treatment for children with attention-deficit/hyperactivity disorder. *Pediatrics*, 118(5):1530–1540, 2006.

[79] T. Surmeli and A. Ertem. QEEG guided neurofeedback therapy in personality disorders: 13 case studies. *Clinical EEG and Neuroscience*, 40(1):5–10, 2009.

[80] D. Taylor, S. H. Tillery, and A. Schwartz. Direct cortical control of 3d neuroprosthetic devices. *Science*, 296(5574):1829–1832, 2002.

[81] C. Velasco. Adaptive interfaces based upon biofeedback sensors. In *International conference on computers for handicapped persons (ICCHP)*, 2004.

[82] E. Vyzas. *Recognition of Emotional and Cognitive States Using Physiological Data*. Phd thesis, Massachusetts Institute of Technology, 1999.

[83] J. Walker. Power spectral frequency and coherence abnormalities in patients with intractable epilepsy and their usefulness in long-term remediation of seizures using neurofeedback. *Clinical EEG and Neuroscience*, 39(4):203–205, 2008.

[84] J. Walker and G. Kozlowski. Neurofeedback treatment of epilepsy. *Child and adolescent psychiatric clinics of North America*, 14(1):163–76, 2005.

[85] K. Wing. Effect of neurofeedback on motor recovery of a patient with brain injury: A case study and its implications for stroke rehabilitation. *Topics in Stroke Rehabilitation*, 8(3):45–53, 2001.

[86] T. Zander, C. Kothe, S. Welke, and M. Roetting. Enhancing human-machine systems with secondary input from passive brain-computer interfaces. In *Proceedings of the 4th International BCI workshop and Training Course, Graz*, 2008.

Error-Related EEG Patterns during Tactile Human-Machine Interaction

Moritz Lehne
Team PhyPA
TU Berlin, Germany
mle@mms.tu-berlin.de

Klas Ihme
Team PhyPA
TU Berlin, Germany
kih@mms.tu-berlin.de

Anne-Marie Brouwer
TNO Human Factors
Soesterberg, The Netherlands
anne-marie.brouwer@tno.nl

Jan B.F. van Erp
TNO Human Factors
Soesterberg, The Netherlands
jan.vanerp@tno.nl

Thorsten O. Zander
Team PhyPA
TU Berlin, Germany
tza@mms.tu-berlin.de

Abstract

Recently, the use of brain-computer interfaces (BCIs) has been extended from active control to passive detection of cognitive user states. These passive BCI systems can be especially useful for automatic error detection in human-machine systems by recording EEG potentials related to human error processing. Up to now, these so-called error potentials have only been observed in the visual and auditory modality. However, new interfaces making use of the tactile sensory modality for conveying information to the user are on the rise. The present study aims at investigating the feasibility of BCI error detection during tactile human-machine interaction. Therefore, an experiment was conducted where EEG was measured while participants interacted with a tactile interface. During this interaction, errors of the user as well as of the interface were induced. It was shown that EEG patterns after erroneous behavior – either of the user or of the interface – significantly differed from patterns after correct responses.

Keywords: passive brain-computer interface, tactile human-machine interaction, automatic error detection, error potential

1. Introduction

Errors occurring during human-machine interaction (HMI) can have a negative effect on the performance of human-machine systems. It is therefore desirable to design interfaces that are able to detect such errors automatically and – if possible – correct them. However, this is not a trivial task, because in most cases the information available to the interface is not sufficient to reliably detect errors. A promising approach to overcome this problem is the use of brain-computer interfaces (BCIs) that are able to directly extract information from the user's brain, thus providing access to cognitive user states that are not observable from the "outside". In the case of error detection, passive BCIs, i.e., BCI systems that do not rely on conscious effort of the user but extract information of the user's brain without disturbing his primary modes of interaction, could take advantage of specific brain states associated with human error processing [16].

Although the cognitive mechanisms underlying human error detection processes are not yet fully understood, various EEG studies (e.g. [5, 7, 8]) have shown that human error detection is associated with typical patterns in the EEG signal, so-called error potentials. In general, these error potentials can be observed whenever the actual outcome of an action does not match the intended outcome. The exact structure of error potentials depends on the type of error. In the context of human-machine interaction, two basic error types can be distinguished. On the one hand, errors may be committed by the interface (e.g. when it is wrongly interpreting a user intention due to restricted information); on the other hand, errors may occur due to erroneous behavior of the user (e.g. when accidentally pressing a wrong button). In the following, the former will be termed *machine errors* while the latter will be referred to as *self-generated errors*.

EEG patterns related to machine errors are composed of a negative component about 200 ms after occurrence of the error and a positive peak after about 300 ms [6, 16]. While for the negative deflection parietal [16] as well as fronto-central [6] regions were reported, the positive peak is strongest at central electrode sites. In the context of BCI, the EEG patterns related to machine errors have been successfully detected on a single-trial basis in an online BCI

978-1-4244-4799-2/09/$25.00

application [16].

EEG patterns reflecting self-generated errors in visual and auditory tasks were for example demonstrated in [7] and [5]. They manifest themselves in a negative deflection termed error-related negativity (ERN or Ne). This negativity peaks approximately 50–100 ms after the erroneous response over fronto-central regions of the scalp with an amplitude of up to 10 μV. The negative peak is followed by a later positive potential that is labeled as Pe and appears between 200–500 ms over the centro-parietal area [5,14]. Like machine errors, self-generated errors could also be detected on a single-trial basis in an offline study [2].

The previous studies investigated errors in the visual and auditory domain. However, tactile interfaces have recently attracted the attention of HMI reseachers [15]. Advantages of these interfaces include their potential to lower cognitive workload in other modalities and their ability to intuitively direct a user's attention (a proverbial "tap-on-the-shoulder"). Because of the increasing importance of tactile interfaces, it is of interest to determine whether error potentials similar to the ones found in the visual and auditory domain can also be observed for tactile stimuli. Tactile stimuli have been shown to elicit P300 event related potentials [9, 13] that can be used in a BCI system [3]. To the authors' knowledge there exists only one study [12] that indicates that self-generated error potentials can be elicited in the tactile domain. In this study, participants performed a time estimation task, in which they received tactile feedback about whether their estimation had been correct or not. Whenever the feedback informed them about an incorrect estimation an error potential similar to the ones observed in the visual and auditory modality occurred.

The present study explores the EEG patterns related to error processing in the tactile modality for both, self-generated and machine errors. We will address the question as to whether cognitive processing of these errors during interaction with a tactile human-machine interface elicits error potentials as described for visual or auditory stimuli.

In order to tackle this question, an EEG experiment was conducted in which participants interacted with a tactile human-machine interface that occasionally committed errors. Additionally, by varying the difficulty of the task, user errors were induced (self-generated errors). Both, averaged and single-trial EEG data was analyzed to investigate the feasibility of automatic error detection during tactile human-machine interaction.

2. Methods

2.1. Participants

Eleven participants (four female) took part in the experiment. All of them were neurologically healthy and had normal or corrected-to-normal vision.

2.2. EEG Recording

EEG activity was recorded at electrodes Fz, FCz, FC1, FC2, Cz, CPz, Pz, and POz of the international 10-20-system mounted on an electrode cap (g.tec medical engineering GmbH). A ground electrode was placed on participants' forehead. Electrodes were referenced to the linked mastoids and impedances of all electrodes were below 5 kΩ. In order to detect eye artifacts, electrooculogram (EOG) was measured via two eye electrodes referenced to each other. Data was recorded with a sampling rate of 256 Hz and filtered with a 50 Hz notch filter, a high-pass filter of 0.1 Hz and a low-pass filter of 60 Hz (USB Biosignal Amplifier, g.tec medical engineering GmbH).

2.3. Task

In one run, participants' task was to move a tactile cursor to the location of a target. To determine the direction in which the cursor would move, visual stimuli indicating either a clockwise or a counterclockwise movement were presented to participants. Participants could either select or reject the direction of movement by pressing an 'accept' or a 'reject' button, respectively. When accepting, the cursor moved to the next location on the tactile display in the designated direction; otherwise it stayed at the current location. A run ended when target and cursor were at the same spot.

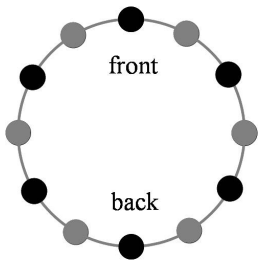


Figure 1. Schematic top-view of the tactor locations at one level of the TNO tactile torso display as used in the experiment. Front indicates position of navel. Possible target positions are colored black. Every tactor could be a cursor position.

2.4. Stimuli

Tactile stimuli. Tactile stimuli were presented via the TNO tactile torso display [15], a wearable vest containing five rows of twelve equally spaced, custom-built tactors; they consisted of plastic cases with a contact area of 1×2 cm, containing motors vibrating at 160 Hz (TNO, The Netherlands, model JHJ-3). The adjustable vest was worn above the clothes and spanned the whole trunk circumference of the participant. In the current study only the central row of tactors was used for stimulus presentation. During one trial, one tactor was target while another one was cursor. Possible target and cursor locations are shown in Figure 1.

The target continuously switched between 100 ms intervals of vibrating and not vibrating. In contrast, the cursor only vibrated once per trial for 400 ms before presentation of the visual stimulus.

Visual stimuli. Visual stimuli were presented on an LCD (Dell 20" flat panel, refresh rate 75 Hz). Every stimulus consisted of eight black or grey arrows arranged in a circle pointing either clockwise or counterclockwise (see Figure 2). The stimuli were 1280×1024 pixels in size. For every trial one of the stimuli was presented with a duration of 200 ms.

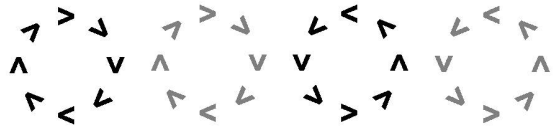


Figure 2. The four different visual stimuli used in the experiment.

2.5. Conditions

Three different experimental conditions were used: *machine error only*, *self-generated error only* and *mixed*. The names were given according to the type of errors that were induced in the respective blocks of the experiment.

In the 'machine error only' condition, the direction of the arrows always corresponded to the movement direction that was to be accepted or rejected, thus making the task easy. However, in 20 % of the cases the cursor moved opposite to the direction intended by the participant. Note that two errors could not occur in a row.

In the 'self-generated error only' condition, the color of the arrows determined whether the cursor movement would be congruent or incongruent with the indicated direction. Arrows in black led to a movement in the direction indicated by the arrows. Contrarily, arrows in grey led to a movement in the opposite direction as indicated. Thus, when a participant wanted the cursor to move in clockwise direction, he or she should press the 'accept' button when presented with black arrows indicating a clockwise direction, whereas he or she should press the 'reject' button when presented with grey clockwise arrows. This more difficult task was used to induce errors in the response of the participants. No machine errors occurred in this condition.

In the mixed condition self-generated errors and machine errors were combined, i.e., the task was hard and machine errors occurred with a probability of 20 %. Table 1 presents an overview of the three conditions.

2.6. Experimental Design

Participants performed two experimental blocks per condition in random order. Each block consisted of six runs. In each run, the target appeared in one of the six possible target locations (see Figure 1). Each target location occurred twice per condition; once the cursor appeared three steps left from the target (i.e., it had to be moved clockwise) and once it appeared three steps right (i.e., it had to be moved counterclockwise). Order of target location was random.

	Condition		
	machine error only	self-generated error only	mixed
task difficulty	easy	hard	hard
machine error rate (in %)	20	0	20

Table 1. Overview of experimental conditions, the corresponding task difficulty and the percentage of machine errors.

2.7. Procedure

Participants were seated comfortably in front of a monitor in a dimly lit, shielded room, wearing the tactile display and the EEG electrode cap. Before the experiment the task was explained to participants. They were told to always press one of the buttons ('accept' or 'reject') and were informed that occasionally the interface may commit errors.

The experiment started with a training session consisting of four runs for both the easy and the hard task. During training, no machine errors occurred. The training lasted about eight minutes. After that, participants performed the six experimental blocks. Between blocks, participants had a short break. One block lasted about six minutes. During the experiment pink noise was presented to participants to mask the noise of the vibrating tactors.

Each run consisted of several trials repeating until the cursor reached the target (for a schematic overview of one trial, see Figure 3). During a trial the target vibrated as described in section 2.4. 1100 ms after the start of a trial, the cursor vibrated for 400 ms. After 800 ms plus a stimulus onset asynchrony (SOA) of 0 to 300 ms, a visual stimulus, randomly picked from the set described above, was presented to participants for 200 ms. Then, a response time of maximal 1500 ms followed, in which participants should press a button. If they failed to press within this time interval, a visual message telling them to react faster was presented for 900 ms, starting 100 ms after the end of the maximal response time and ending 100 ms before the start of the next cursor vibration. If a button was pressed in time, a blank screen was displayed instead. Upon reaching the target, a

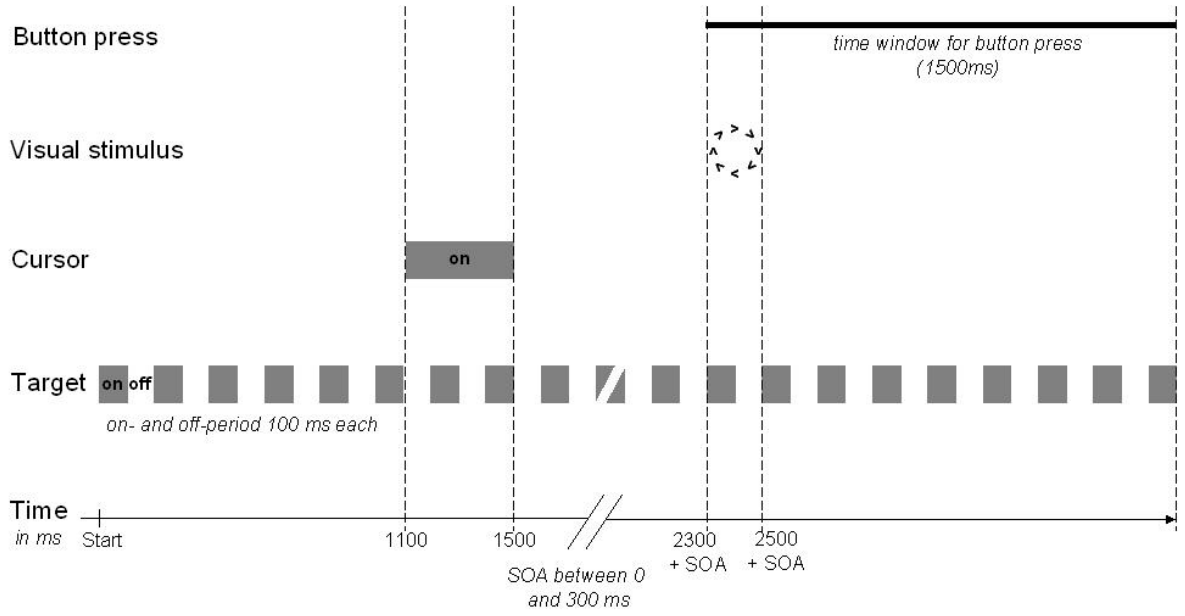


Figure 3. One experimental trial.

visual message 'Target found' was displayed.

2.8. Analysis

Artifact removal. To clean EEG data from artifacts, sections containing data that could neither be attributed to brain activity nor to eye movements were rejected manually. None of the experimental trials were affected by this. Afterwards, an independent component analysis (ICA; as described in [11]) was computed with the open source toolbox EEGLAB [4] in order to identify and remove components reflecting eye activity. For this, the eight EEG channels plus the two eye electrodes were used as input. The ICA delivered nine independent sources. To identify components reflecting eye artifacts, we used the following three criteria: (1) the dipole of components reflecting eye activity has a fronto-lateral distribution, (2) the power spectra of eye artifacts do neither show a clear peak in the alpha range nor a sharp decrease of power with increasing frequency, and (3) the peaks of the component occur at the same latencies as the ones from the EOGs. Following these criteria, between one and two eye artifact components were identified per participant and removed accordingly. Data were transformed back from the remaining components to the channel representation. Analysis of event-related potentials (ERPs) was done on these artifact-free data.

ERP analysis. For ERP analysis, EEG from error trials was compared to the EEG from correct trials. For machine errors, the ERPs locked to the presentation of the erroneous cursor move were compared to the ERPs locked to the presentation of the correct cursor move. Cursor presentations after erroneous button presses were excluded. Machine errors during the last step of the run were discarded in order to avoid effects of the visual message 'Target found'. For self-generated errors, the ERPs locked to the erroneous button press were compared to the ERPs locked to correct button presses.

For statistical analysis the error negativity and error positivity as described in [16] and [6] for machine errors were identified in the grand average (i.e., the averaged ERP over all participants) and time windows representing these peaks were defined. In the self-generated error condition, the same was done for the Ne and Pe. Within these time windows the mean of the EEG signal of each trial was calculated. T-tests were used to compare the distributions of these mean values. Significance levels for the t-tests were lowered according to the Bonferroni method.

For machine errors, analysis was performed for electrodes Fz, FCz, Cz, Pz, and POz. The first three were chosen because of successful classification on similar channels in [6], Pz and POz were added because of the classification results in [16].

For self-generated errors, comparisons were carried out for electrodes FCz, Cz and CPz, the regions where Ne and Pe are supposed to peak according to [6].

In order to maximize the amount of trials for each error type, analysis was conducted on the pooled data from the different conditions. This means that machine error trials were extracted from the conditions 'machine error only' and 'mixed'. Likewise, self-generated error trials were taken

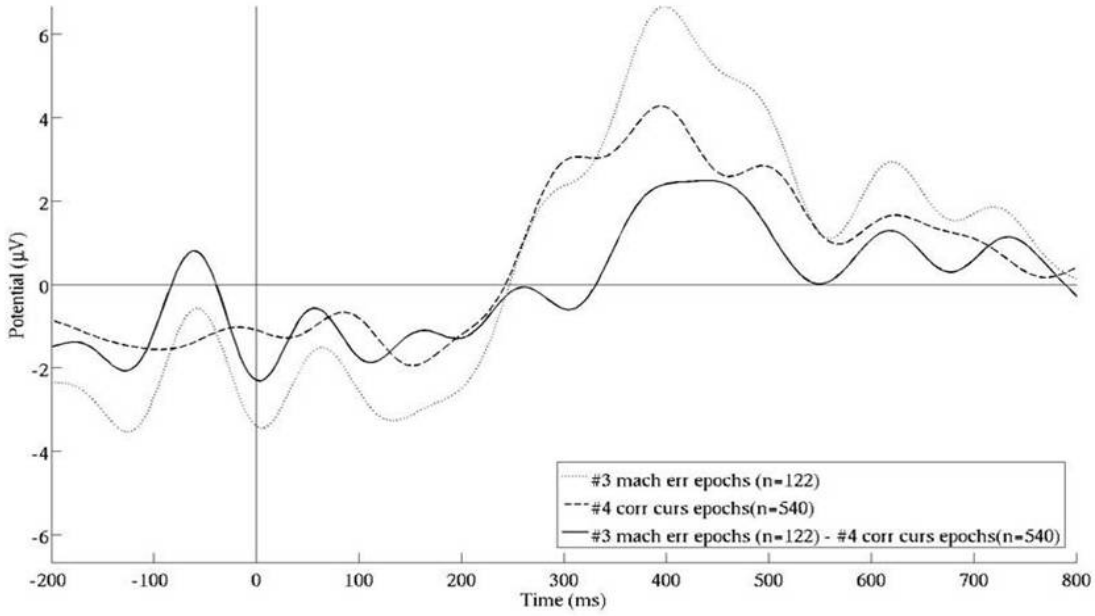


Figure 4. Averaged EEG data time-locked to the beginning of the cursor vibration pooled from conditions 'machine error' and 'mixed' at electrode CPz. Averages over trials with a machine error (dotted line, n = 122) and trials with a correct cursor movement (dashed line, n = 540) are shown separately. The solid line shows the difference between machine errors and correct cursor movements.

	Fz		FCz		Cz	
	140-210 ms	360-500 ms	140-210 ms	360-500 ms	140-210 ms	360-500 ms
p	0.75	**0.002***	0.69	**0.001***	0.60	0.007
df	660	660	660	660	660	660
t	-0.3	3.07	-0.41	3.32	-0.53	2.69

	Pz		POz	
	140-210 ms	360-500 ms	140-210 ms	360-500 ms
p	0.10	**0.004***	0.09	0.04
df	660	660	660	660
t	-1.62	2.92	-1.68	2.03

Table 2. Results of the t-tests for the comparison 'machine error' versus 'correct cursor movement' for the five electrodes and the different time windows 140-210 ms and 360-500 ms. Significant differences are written in bold.

from the 'self-generated error only' and the 'mixed' condition.

2.9. Classification

Though providing a valuable tool for the investigation of event-related potentials in human error processing, statistical analysis of averaged data is not sufficient for BCI applications. A successful BCI for error detection essentially depends on differences in the EEG data that can be detected on a single-trial basis.

To investigate whether single-trial detection of errors is possible, the performance of a BCI classifier on the data was evaluated. Therefore, epochs for correct and error trials were extracted from the raw EEG signal. Based on the results of the analysis of averaged ERP data, time windows of 150 ms subdivided into three parts of 50 ms length were defined. The resulting 24-dimensional feature space (3 time window subparts $\times$ 8 EEG channels) was then used for training a linear classifier (regularized linear discriminant analysis) to distinguish between correct and error trials for machine and self-generated errors. These feature extraction and classification methods were chosen because of their performance in a study benchmarking common BCI algorithms [10]. To avoid overfitting of the classifier, evalu-

Figure 5. Averaged EEG data time-locked to the button presses pooled from conditions 'self-generated error' and 'mixed' at electrode FCz. Averages over trials with a self-generated error (dotted line, n = 237) and trials with a correct cursor movement (dashed line, n = 1211) are shown separately. The solid line shows the difference between self-generated errors and correct button presses.

	FCz		Cz		CPz	
	40-110 ms	210-310 ms	40-110 ms	210-310 ms	40-110 ms	210-310 ms
p	$<$ **0.001****	$<$ **0.001****	$<$ **0.001****	$<$ **0.001****	$<$ **0.001****	$<$ **0.001****
df	1446	1446	1446	1446	1446	1446
t	-7.51	4.71	-7.92	4.37	-7.46	4.12

Table 3. Results of the t-tests for the comparison 'self-generated error' versus 'correct button press' for the three electrodes and the different time windows Ne (40-110 ms) and Pe (210-310 ms). Significant differences are written in bold.

ation of the classification process was done using a tenfold cross-validation.

3. Results

3.1. Descriptive Results

Machine errors. On average 11.1 (standard deviation 4.4) machine errors per participant were taken into account for analysis (machine errors on the last step and after self-generated errors were discarded).

Figure 4 shows an example ERP at electrode CPz averaged over all subjects. Visual inspection of the ERP plots did not reveal a clear negative component. A positive deflection was visible around 360 to 500 ms.

Self-generated errors. The mean amount of self-generated errors was 27.8 (standard deviation 15.2). Figure 5 shows an ERP at channel FCz. A negative component peaked between 40 and 110 ms after the button press, while a smaller positive component occurred at a time window from 210 to 310 ms. Additionally, there was a large difference between the trials with erroneous and correct button presses during the 80 ms before the button press.

3.2. ERP analysis

Machine errors. Table 2 displays the results of the t-tests for the machine errors for different electrode sites and time windows. Although no clear negative component was observed, t-tests were calculated for the time window in which the component was supposed to occur (140–210ms).

subject	# error trials	# correct trials	classification accuracy in % time window 360-510 ms		
			error trials	correct trials	overall
1	19	57	69	69	69
2	11	47	63	75	69
3	12	51	60	63	62
4	7	54	44	81	77
5	5	56	18	85	80
6	10	56	50	80	74
7	6	40	50	71	68
8	14	51	29	56	50
9	12	59	41	55	53
10	9	55	67	82	80
11	17	66	52	72	68
mean	11.1	53.8	49.4	71.7	68.2

Table 4. Overall and per-class classification accuracies for machine errors in the time window 360-510 ms after onset of stimulus presentation.

Since ten t-tests were calculated, the significance level was lowered to 0.005 according to the Bonferroni method. Differences in the early time window reflecting the negative component are not significant. In the time window reflecting the positive component, EEG differences between the error trials and the correct cursor movements are significant at all electrode sites except Cz and POz.

Self-generated errors. The results of the t-tests for the self-generated errors for different electrode sites and time windows are shown in Table 3. Again, hypotheses were tested against a significance level of 0.005. All comparisons in both of these time windows yielded significant differences. Likewise, for the time window before the button press (-80 to 0 ms) all t-tests showed significant results (FCz: $t(1446) = -4.62$, $p < 0.001$**, Cz: $t(1446) = -5.48$, $p < 0.001$**, CPz: $t(1446) = -5.39$, $p < 0.001$**).

3.3. Classification

Table 4 displays the classification results for the individual participants in the machine error condition for the time window 360 to 510 ms after onset of the tactile stimulus. Overall classification accuracies as well as accuracies for individual classes are reported. Mean classification rate was 68.2 %, and none of individual classification accuracies was below 50 %. However, accuracy on error trials was lower (49.4 %) than on correct trials (71.7 %). For two participants (8 and 9) classification accuracy was considerably lower than average. Table 5 presents the results for the human error condition. Mean classification accuracy was 70.4 % (error: 52.1 %, correct: 72.6 %) for the time window 0 to 150 ms after the button press and 68.5 % (error: 45.8 %, correct: 71.4 %) for the time window 170 to 320 ms.

4. Discussion

4.1. ERP analysis

Machine errors. The comparison between machine error trials and trials with correct cursor movements yielded significant differences that are revealed at electrode sites Fz, FCz and Pz. To our knowledge, this is the first study showing that error potentials can be elicited by machine errors in a tactile task.

The main difference between error and correct trials was found in a positive deflection in the range of 360 to 500 ms. Although peaking somewhat later, this component seems to reflect the positive peak reported in [16] and [6] for the visual domain. The increased latency of this positive deflection might be caused by longer processing times in the somatosensory modality. It was shown for example in an oddball paradigm with visual, auditory and tactile stimulation that P300s elicited by tactile and auditory stimuli peaked 200 ms later than for visual stimuli [1]. Contrarily, Miltner, Braun and Coles [12], reported an earlier peak of the error-related negativity after participants received feedback in the somatosensory domain. It might therefore be possible that other stimulus and task characteristics than stimulus modality are responsible for the variability in latencies.

Theoretically, the positive peak could be explained by the fact that the error was accompanied by a change in cursor direction. This produced a more or less rare event on a lower level of cognitive processing than a (rare) error. Thus, part of the effect could have been caused by processing of a rare event which is comparable to an oddball paradigm eliciting a P300. However, one should note that the cursor started only three steps away from the target, and the direction of cursor movement was randomly varied such that overall, the clockwise or counterclockwise direction hap-

subject	# error trials	# correct trials	classification accuracy in % time window 0-150 ms			time window 170-320 ms		
			error trials	correct trials	overall	error trials	correct trials	overall
1	39	182	59	69	67	62	74	72
2	9	151	67	83	82	11	77	73
3	41	175	59	68	67	54	70	65
4	9	160	45	81	80	33	88	75
5	46	174	41	61	57	44	59	55
6	21	171	43	80	75	47	62	66
7	8	136	50	83	81	25	82	79
8	18	164	39	69	66	61	74	73
9	40	177	43	67	63	62	68	67
10	30	167	63	65	65	50	65	63
11	45	219	64	73	71	55	66	65
mean	27.8	170.5	52.1	72.6	70.4	45.8	71.4	68.5

Table 5. Overall and per-class classification accuracy for self-generated errors in the time windows 0-150 ms and 170-320 ms after button press.

pened equally often. Also, Ferrez and Millán [6] who used a similar protocol, observed the positive peak with error rates of 20 % as well as 50 %. Nevertheless, in upcoming studies the experimental paradigm should be adapted so that possible P300 effects are excluded as far as possible. A promising approach might be to use the experimental paradigm suggested in [16], which rules out direction effects.

In the present study, we did not find a negative component around 200 ms after wrong cursor movements as described for machine errors in the visual domain [6]. This might be due to the small number of trials. Increasing the number of trials in future studies might deliver clearer components. Another explanation for the missing negative component is that the probability of machine errors was too high. Previous research found that the less likely the wrong feedback, the more prominent the amplitude of the negativity [14]. However, [16] and [6] detected this component with the same or an even higher error rate than we used here.

Self-generated errors. The results observed here are well in line with the results reported in [2], i.e., both, an early negative deflection ascribable to the Ne and a late positive deflection that can be attributed to the Pe, were observed. So, it seems that self-generated errors committed in a task where feedback is given in the tactile domain can elicit error potentials in a similar manner as visual or auditory tasks. This supports a generic and modality independent error detection system in the human brain as proposed in [12].

Additionally, significant differences were found in a time window from 80 ms before the button press to the beginning of the button press. This might be because participants realize their mistake even before pressing the button, but cannot suppress the motor action anymore.

4.2. Classification

Mean individual classification accuracies were all above 50 %. However, the individual classification accuracies for the two classes were very different. Especially for two participants, accuracies for both classes were lower than average. This might be attributed to the fact that the time course of the error potential deviated from the time window chosen for classification. Defining time windows for individual participants may thus improve classification.

From the point of view of usefulness during human-machine interaction, it is unlikely that performance can be enhanced by our system, since in most cases there are more false alarms and missed errors than correctly identified errors. This could be due to two different reasons. First of all, error potentials might not be elicited in tactile tasks. However, this is not supported by the results of the ERP analysis which showed significant differences between error trials and correct trials. So it seems more probable that the low performance of the classifier is due to the small amount of error trials used in this study, which makes it hard for the classifier to estimate the underlying distributions. The question of whether single-trial classification of tactile error potentials is possible should therefore be further investigated on a larger set of trials. Furthermore performance should be evaluated in an online experiment, as already done in the visual domain by Zander et al. [16].

5. Conclusion

In the present study, it was shown that errors occurring during tactile human-machine interaction give rise to specific patterns in the EEG signal, which differ in structure depending on the type of error that is induced. While machine errors are mainly characterized by a positive deflec-

tion peaking at about 400 ms after the occurrence of an error, self-generated errors generally manifest themselves in a prominent negativity at earlier times of the ERP followed by a later positivity. The effects were clearly observable in averaged data. Furthermore, single-trial classification accuracies were higher than 50 % for all subjects. However, in order to reliably evaluate the performance of a BCI classifier for error detection, the results of this study have to be re-evaluated on a larger amount of data. For future studies it would also be interesting to investigate the spatiotemporal patterns of error potentials more thoroughly. Therefore recording more EEG channels might provide further insights.

6. Acknowledgements

We thank Antoon Wennemers for his great help in programming the experiment and gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

[1] F. Aloise, I. Lasorsa, F. Schettini, A. Brouwer, D. Mattia, F. Babiloni, S. Salinari, M. Marciani, and F. Cincotti. Multimodal stimulation for a p300-based BCI. *International Journal of Bioelectromagnetism*, 9(3):128–130, 2007.

[2] B. Blankertz, C. Schäfer, G. Dornhege, and G. Curio. Single trial detection of EEG error potentials: A tool for increasing BCI transmission rates. In *Artificial Neural Networks ICANN 2002*, pages 1137–1143, 2002.

[3] A. Brouwer and J. B. F. van Erp. A tactile p300 BCI and the optimal number of tactors: effects of target probability and discriminability. In *Proc. of the 4th Int. BCI Workshop & Training Course*, Graz, Austria, 2008. Graz University of Technology Publishing House.

[4] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[5] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein. ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, 51(2-3):87–107, 2000.

[6] P. Ferrez and J. del R. Millan. Error-Related EEG potentials generated during simulated BrainComputer interaction. *Biomedical Engineering, IEEE Transactions on*, 55(3):923–929, 2008.

[7] W. J. Gehring, B. Goss, M. G. H. Coles, D. E. Meyer, and E. Donchin. A neural system for error detection and compensation. *Psychological Science*, 4(6):385–390, 1993.

[8] C. B. Holroyd and M. G. H. Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4):679–709, 2002.

[9] J. Ito and N. Takamatsu. Somatosensory event-related potentials in healthy subjects: single trials analysis and averages of reaction time terciles. *Journal of Psychophysiology*, 11:2–11, 1997.

[10] C. Kothe and T. Zander. Benchmarking common BCI algorithms for fast-paced HMS applications. In *Proc. of the 4th Int. BCI Workshop & Training Course*, Graz, Austria, 2008. Graz University of Technology Publishing House.

[11] S. Makeig, A. J. Bell, T. P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, pages 145–151, 1996.

[12] W. H. R. Miltner, C. H. Braun, and M. G. H. Coles. Event-Related brain potentials following incorrect feedback in a Time-Estimation task: Evidence for a Generic neural system for error detection. *Journal of Cognitive Neuroscience*, 9(6):788–798, 1997.

[13] Y. Nakajima and N. Imamura. Relationships between attention effects and intensity effects on the cognitive n140 and p300 components of somatosensory ERPs. *Clinical Neurophysiology*, 111(10):1711–1718, 2000.

[14] S. Nieuwenhuis, K. R. Ridderinkhof, J. Blom, G. P. Band, and A. Kok. Error-Related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38(05):752–760, 2001.

[15] J. B. F. van Erp. *Tactile displays for navigation and orientation: perception and behaviour*. PhD thesis, University of Utrecht, the Netherlands, 2007.

[16] T. O. Zander, C. Kothe, S. Welke, and M. Rötting. Enhancing human-machine systems with secondary input from passive brain-computer interfaces. In *Proc. of the 4th Int. BCI Workshop & Training Course*, Graz, Austria, 2008. Graz University of Technology Publishing House.

Sparse matrix factorization for Brain Computer Interfaces

Alberto Llera Arenas
Donders Institute/Biophysics Department
Radboud University Nijmegen
The Netherlands.
a.llera@donders.ru.nl

Vicenç Gómez
v.gomez@science.ru.nl

Hilbert J. Kappen
B.Kappen@science.ru.nl

Abstract

We present a novel sparse dimensionality reduction approach to reconstruct biological signals for brain computer interfaces (BCI). The proposed technique may be used in the design of an adaptive Brain Computer Interface which uses interaction error potentials.

1. Introduction

Interaction error potentials (IEP) are potentials detected in the recorded EEG of a subject controlling a device, just after the device performs an error. The error is the difference between the result of the action that the subject expected, based on his/her action, and the actual outcome.

Since the 1990's there has been many studies related to the presence of error potentials. They can be classified as follows: the response error potential [3] found in speeded reaction tasks; the feedback error potential [8] which appears in reinforcement learning tasks; the observation error potential [12] and finally, the IEP, which can be detected in a Brain Computer Interface (BCI) context [4].

The precise detection of an IEP after the BCI makes a classification error can help us to construct a more robust BCI, by either correcting the BCI output directly, or more interestingly, by adapting the BCI classifier so that it is less likely to make a similar mistake in the future. This idea is illustrated in Figure 1.

From EEG studies it is well known [3, 4, 8, 12] that the error (as we introduced above) is usually followed by what is called event-related negativity (ERN) which is found in the α-band in fronto-central channels. More recently, an MEG study [7] on the detection of error fields in MEG has shown an increase in the frontal μ-power and a decrease in the posterior α and central β-power.

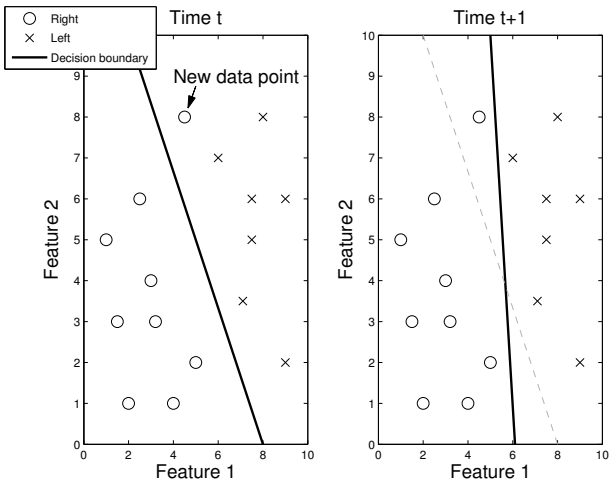


Figure 1. Illustration of an adaptive BCI for a binary task. Each point is labeled with the movement (class) that was intended by the subject (left or right) and denotes a brain state encoded using two features. Bold line indicates the decision boundary of the BCI classifier. **(Left)** A new point is misclassified. The IEP recognized by the BCI provides a mechanism to detect the misclassification. **(Right)** The decision boundary is changed and the BCI is adapted during performance.

The application of the IEP to BCI [4] requires its reliable detection. The IEP may in principle be localized in various channels, various frequency bands, and may be subject dependent.

In this paper, we propose a novel dimensionality reduction approach which can be used to analyse the IEP. We propose a sparse version of singular value decomposition (SVD) that describes the high dimensional signal as a sum of a small number of sparse templates that change through time. The sparsity means that the number of channels that are used in each template is small and it will greatly im-

978-1-4244-4799-2/09/$25.00

prove the interpretability of our findings. Our approach is then related to works which rely on signal decomposition using different spatio-temporal features [6, 10] and opens new doors on how to classify the interaction error fields (IEF, which are the MEG equivalent of the IEP) since we do not need to focus in just a few electrodes, but we may use all the electrodes to increase the quality of the classification.

Our approach is presented without any preselected frequency band for detection of IEF, since the aim of this work is only to present a new method and is not specially focused on solving the IEF classification. However, it can be applied to any specific frequency band that previous knowledge might indicate is the most relevant for a particular problem.

2. Experimental setup

We describe now the experimental framework we used in our data acquisition. The main goal of this experimental design is to gain insight into how error signals are encoded in the brain. Up to now, we gathered measurements from two subjects. Each subject performed 6 sessions composed of 84 trials with a minute between two sessions. We plan to acquire data from 25 more subjects.

All the data used during this work was collected using an MEG system with 275 channels from which 273 were in use. EOG and ECG were also recorded and trials with ocular or muscular artifacts were removed from the data using an automatic routine.

The experiment is designed as follows:

1. First, two squares and a fixation cross appear in the screen.

2. After 300 ms, the fixation cross becomes an arrow (pointing to left or right). The subject is instructed to direct the attention to the direction pointed by the arrow points *while keeping the sight in the center of the screen*.

3. After 2000 ms, the arrow disappears and is replaced with a text indicating the decision of the device (*right* or *left*). This lasts for 1000 ms, and it is the period of main interest.

4. Finally, the text disappears and the two squares remain in the screen for 1000 ms before the new trial starts.

Note that subjects are instructed to control the device using directed (or *covert*) attention, a well known paradigm for BCI control, based on the lateralization of the power on the α band in the posterior channels [11]. However, we could also have used another paradigm such as, for instance, motor imagery, without any change in our protocol.

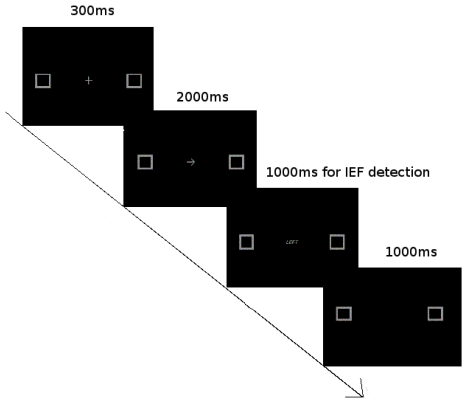


Figure 2. Experimental protocol.

In this preliminary setup, to focus in the goal of error detection, the device returns automatically a random 20% of error responses. We labeled as *error* trials those with the wrong feedback (when the text does not correspond to the direction pointed by the arrow) and *correct* trials otherwise.

The length of the trials was reduced to 1800 ms. For that we selected the full period for IEF detection (1000 ms) plus 800 ms of the arrow. The recording sampling rate was 1200 Hz which gave us a total of 2160 time points per trial. This means that our data matrix for a single trial has size $n \times t$, where $n = 273$ and $t = 2160$.

3. Theoretical framework

In this section we present our method to obtain a reconstruction of the data using a reduced and sparse set of features. First, we describe how we perform dimensionality reduction and then we focus on sparsity.

3.1. Matrix Factorization and Dimensionality Reduction

Lets assume that we have a data matrix $Y \in \mathcal{M}_{n \times t}$ where n and t indicate number of channels and time-steps respectively. When facing the problem of matrix factorization in a general setting, our goal is to find two matrices F and G that minimize

$$||FG - Y||_2^2. \tag{1}$$

where $||FG - Y||_2$ is the Frobenius norm of the matrix $FG - Y$. This can be seen as constructing a basis matrix F for which the coefficients for the data are in matrix G.

A common first step when classifying data is to reduce the effect of the noise and use the most informative features. This is usually done using dimensionality reduction techniques. In our case, we retain the most informative k basis vectors and discard the rest.

In this general setting, we see that for any given $k \in \mathbb{N}$ we can find matrices $F \in \mathcal{M}_{n\times k}$ and $G \in \mathcal{M}_{k\times t}$ that minimize expression (1). We are interested in the case where $k \ll n$. Here appears the model selection problem, or how to select the parameter k.

For $k = n$, the singular value decomposition (SVD) can be used to factorize Y and obtain three matrices: $U \in \mathcal{M}_{n\times n}$, a diagonal matrix $S \in \mathcal{M}_{n\times t}$ and $V' \in \mathcal{M}_{t\times t}$ such that

$$Y = USV^*. \tag{2}$$

where * denotes conjugate transpose of a matrix, and the singular values of Y are sorted by their absolute value in descending order along the diagonal of S. If we define $\mathbf{F} = U$ and $\mathbf{G} = SV^*$, such a factorization corresponds to the minimization of (1) for the case of $k = n$.

For $k \ll n$, we define the matrix F considering only the first k columns of $\mathbf{F}$ and equivalently, G considering the first k rows of $\mathbf{G}$. Hence, FG is an approximation of Y, which becomes more accurate as k increases.

In this work, we use the *Akaike information criterion* (AIC) [1] to select the value of k. In our particular case, under the assumption that errors are normally distributed, the AIC selects the k which minimizes

$$||FG - Y||_2^2 + k(n + t). \tag{3}$$

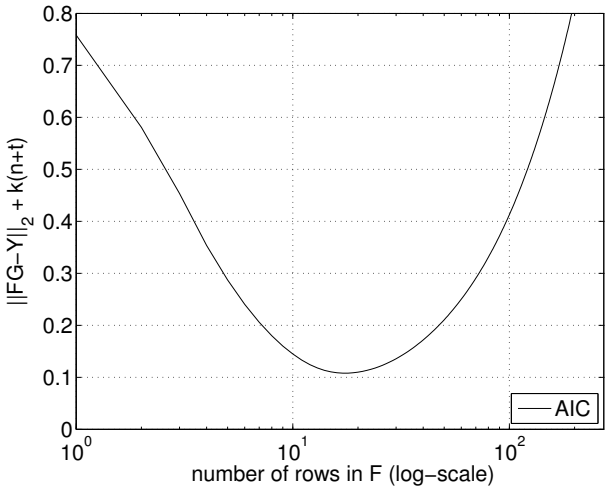


Figure 3. The Akaike information criterion (AIC) for model selection is used to select the number of features k in our approach.

Figure 3 shows the AIC for different trials corresponding to the experiment described in section 2. From now on we can assume that k is fixed.

3.2. Sparsity

Up to now we have described how to represent the matrix Y using a *reduced* basis F of k vectors. Each of the k vectors can be considered as a feature composed of a mixture of different channels. To reconstruct the original signal over time these features are weighted by the corresponding coefficients in G.

In this section we explain how to make the basis F sparse. Enforcing sparsity in F will result in features composed of a reduced number of channels thus providing a more compact and structured representation of the data and consequently, increasing the interpretability of the recovered signal.

A natural method to obtain a sparse F is an extension to matrices of the ℓ_1-norm regularized least squares method [2]. Given the data Y, and assuming an initial G fixed, we are interested in the F which minimizes

$$||FG - Y||_2^2 + \lambda||F||_1, \tag{4}$$

where the $||F||_1$ is the sum of the absolute values of the elements in the matrix F.

Instead of minimizing Equation (4) directly, we make use of an extension of the algorithm described in [5]. Given a matrix A and a vector y, [5] describes an interior-point method for solving x which minimizes:

$$||Ax - y||_2^2 + \lambda||x||_1. \tag{5}$$

Note first that the minimization of (4) is equivalent to the minimization of

$$||G^{\mathrm{T}}F^{\mathrm{T}} - Y^{\mathrm{T}}||_2^2 + \lambda||F||_1. \tag{6}$$

Thus [5] gives a solution to our problem for $n = 1$.

Now denote the s-th column of Y^{T} by Y_s^{T}. Using [5] we can also find a solution to

$$||G^{\mathrm{T}}x - Y_s^{\mathrm{T}}||_2^2 + \lambda||x||_1, \tag{7}$$

where x is exactly the s-th column of F^{T}. Repeating this procedure for every $s \in \{1 \dots n\}$ we can find F, a solution of (6) and consequently of (4). In other words, we have expressed the global minimization (4) as n independent minimizations of the form (7), one for each channel.

Parameter λ plays the role of a trade-off between sparsity and quality of the reconstruction. On one hand, for a small λ, the quality of the reconstructed signals will be high. However, F will less sparse. On the other hand, a large λ will result in a very sparse F, but in poor approximations of the original signals.

After having defined a procedure to find a reduced and sparse basis F, we can find a new G which minimizes Equation (1). Since (1) is a differentiable quadratic form in G, the solution can be found analytically and we can write the optimal G in closed form:

$$G = (F^{\mathrm{T}}F)^{-1}(F^{\mathrm{T}}Y). \tag{8}$$

Note that the inverse $(F^{\mathrm{T}}F)^{-1}$ is only defined when $\text{rank}(F) = k$, and this is not generally guaranteed. In particular, the more sparse F is, the more likely is that $\text{rank}(F) < k$. This means that there exists a maximal λ which limits the level of sparsity that can be achieved by our method. In practice, this limitation does not restrict our method, as we will show in the next sections.

4. Algorithms for Sparse matrix factorization

After introducing the theoretical building blocks of our approach, we present two possible algorithms. Both algorithms take as input the BCI data Y, the regularization parameter λ and the desired sparsity of the solution (number of zero entries in F).

Algorithm 1 applies SVD to the original signal Y and then uses AIC (see Section 3.1) to select k. This results in a matrix G with k rows which is used in the ℓ_1-norm minimization (step 4 of Algorithm 1) to find the sparse basis $F^\star$. After the minimization, some of the entries in $F^\star$ are very small in absolute value. We set the required entries to zero as long as the matrix $F^\star$ has full rank (in practice, we always found full rank matrices even using 50% of sparsity).

Algorithm 1

Require: x (number of zeros in F), λ and matrix Y
1: $\mathbf{G} \leftarrow SVD(Y)$.
2: $k \leftarrow AIC$.
3: $G \leftarrow$ select k rows of $\mathbf{G}$.
4: $F^\star \leftarrow \text{argmin}_{F'} ||F'G - Y||_2^2 + \lambda ||F'||_1$.
5: **repeat**
6: $\quad (i,j) \leftarrow$ find smallest non-zero absolute value $F^\star$
7: $\quad F^\star(i,j) := 0$
8: **until** $F^\star$ has x zeros or $\text{rank}(F) < k$.
9: $G^\star \leftarrow \text{argmin}_G ||F^\star G - Y||_2^2$
10: **return** $F^\star, G^\star$

Figure 4 shows the behavior of the algorithm for three different values of λ as a function of the number of zeros. As can be seen, the larger the λ, the more sparse can F be made without increasing significantly the error. Note, however, that for small λ, the initial errors (those corresponding to non-sparse solutions) are smaller than for large λ.

The interplay between λ and the level of sparsity suggests a modification of the algorithm in which the matrix F resulting from SVD, instead of $F^\star$, is used as a final basis. The latter is used only to select which entries of F must be zero. Algorithm 2 describes this alternative approach.

Figure 5 shows a comparison of both methods for a fixed $\lambda = 300$ as a function of the number of zero entries. As can be seen, the alternative algorithm performs better than the previous one as long as the solution is not very sparse.

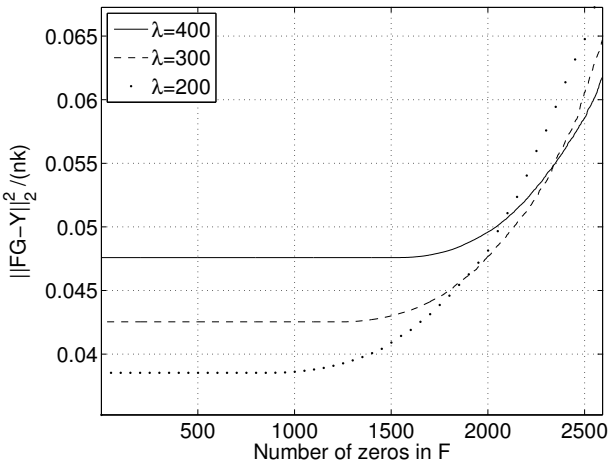


Figure 4. Performance of Algorithm 1 from one random trial. k is fixed to 19 using AIC and = {200; 300; 400}.

Algorithm 2

Require: x (number of zeros in F), λ and matrix Y
1: $\mathbf{F}, \mathbf{G} \leftarrow SVD(Y)$.
2: $k \leftarrow AIC$.
3: $F, G \leftarrow$ select k cols. and rows from $\mathbf{F}, \mathbf{G}$ respectively.
4: $F^\star \leftarrow \text{argmin}_{F'} ||F'G - Y||_2^2 + \lambda ||F'||_1$
5: **repeat**
6: $\quad (i,j) \leftarrow$ find smallest non-zero absolute value $F^\star$
7: $\quad F(i,j) := 0$
8: **until** F has x zeros or $rank(F) < k$.
9: $G^\star \leftarrow \text{argmin}_G ||FG - Y||_2^2$
10: **return** $F, G^\star$

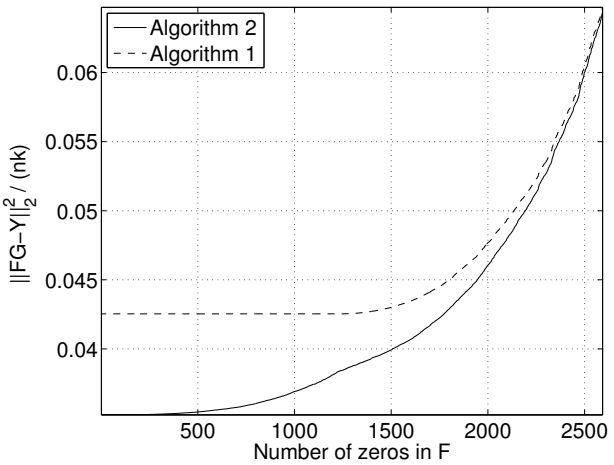


Figure 5. Performance of Algorithm 2 (solid) versus Algorithm 1 (dashed) from one random trial. k = 19 and = 300.

4.1. Choosing the regularization parameter λ

Given a level of sparsity, is there a λ for which the error is minimal? If this is the case, we could choose automatically

the λ provided the number of zero entries in the matrix F.

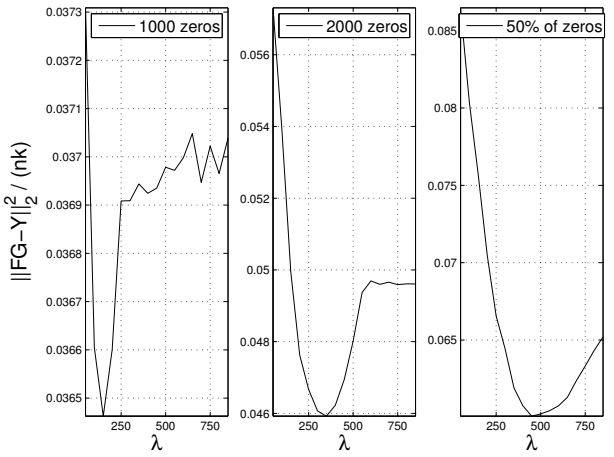


Figure 6. Performance of Algorithm 2 as a function of for different levels of sparsity. Results are equivalent if we consider all the trials.

Figure 6 shows the performance of Algorithm 2 as a function of λ for different levels of sparsity: 1000, 2000 and 2594 (50% of the entries). It shows that there exists a optimal value of λ for any level of sparsity. This optimal value could be easily found, for instance, using line search.

For both algorithms we found that the optimal λ, as well as the error, are larger as we increase the level of sparsity.

4.2. Why not make sparse the SVD directly?

Another way to look at the problem would be to simply make zeros the positions of smallest absolute values of F, and then updating G using (8). In Figure 7 we show that this is not a good strategy. As we can see, the error of Algorithm 2 with $\lambda = 300$ is *always* smaller than this alternative approach, regardless of the level of sparsity, showing the advantage of using the ℓ_1-norm minimization.

This can also be viewed from the perspective that the regularization term used in step 4 of Algorithm 2 has by definition the property to produce parameter shrinkage in the least relevant directions of the data.

5. Results: Sparse reconstruction of signals

In this section we illustrate with an example the quality of the reconstruction made by our method. We show results for the MEG signal acquired according the experimental procedure described in Section 2.

Step 2 of Algorithm 2 gives $k = 19$. Since the MEG system has 273 active channels, this result in a matrix $F \in \mathcal{M}_{273\times 19}$, so the matrix F has a total of 5187 elements. For this example we will require Algorithm 2 to make 2000 zeros in F. For this level of sparsity, we selected $\lambda = 300$.

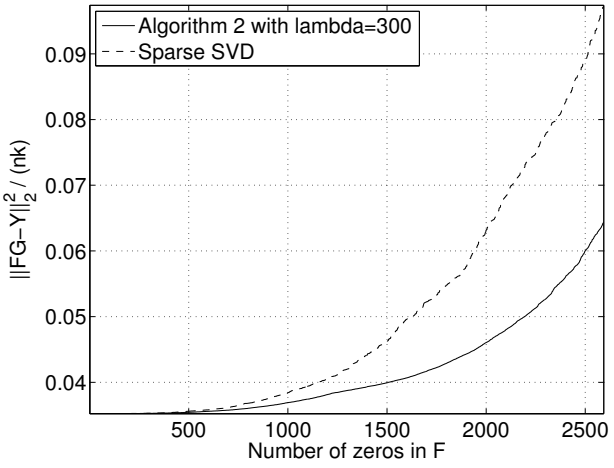


Figure 7. Performance of Algorithm 2 (solid) versus sparsifying the initial SVD (dashed). k = 19 and = 300.

As expected, we observe that columns of F associated with the most relevant features (leftmost columns) are less sparse than the rightmost columns, However, it is not the case that a column becomes totally zero, which would indicate that $\text{rank}(F) < k$.

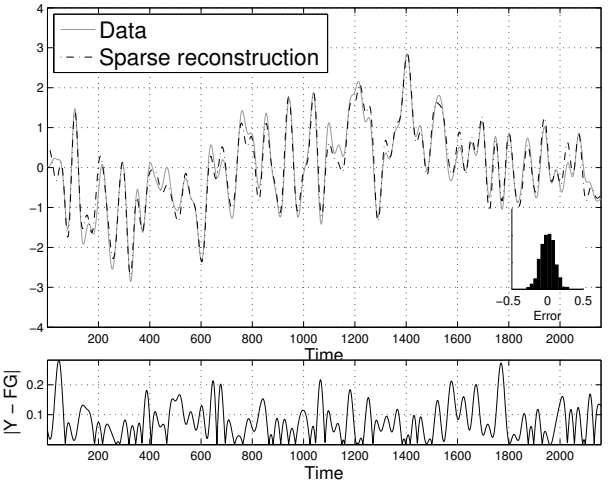


Figure 8. Example of signal reconstruction for one channel and one trial selected randomly. (**Top**) Original signal (grey solid) and the approximation (black dashed-dotted) over time. The approximation was calculated using Algorithm 2 with k = 19, = 300 and 2000 zeros in F. The inset shows an histogram of the residuals, which look normally distributed. (**Bottom**) Residuals as a function of time.

Figure 8 illustrates the reconstruction obtained from a random channel (random row of Y) in one trial using Algorithm 2. For this particular channel there are 7 zeros out of the 19 elements in the respective row of F. The sparsity of the whole matrix F is 38%, whereas the selected channel

appears as irrelevant in 37% of the features. As can be seen from the figure, the reconstruction is very accurate.

5.1. Discussion and ongoing research

We have developed a method to decompose a space/time signal into a small set of features and shown its applicability in MEG signal reconstruction. The method not only leads to a more understandable signal but, more importantly, is also appropriate to be used in a BCI setup, such as the one presented in Section 1, where the reconstructed signal is used in the classification of IEP. This is our current direction of research.

We devise some possibilities to improve/extend the proposed method. First, since the role of the regularizer in our algorithms is just that to select which positions in F should be zero, we might get similar results by using the *Tikhonov* regularization, also known as ℓ_2-regularized least squares [9]. This approach would be much more efficient in computational terms since the regularization becomes a a quadratic differentiable form which therefore has an analytic solution. We have promising preliminary results in this direction.

Another extension is to perform the analysis into the frequency domain, more often used in BCI. Notice that the method can be easily adapted to this case: first, the source data Y would be transformed using a selected band of frequencies (low frequencies are more convenient in our paradigm) and then our sparse factorization would be applied to the transformed data. The resulting basis would constitute a set of spectral features which change over time, the analogous counterpart to our original features.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] S. Boyd and L. Vandenberghe. Effects of error in choice reaction tasks on the ERP under focused and divided attention. In *Convex Optimization*, pages 308–310. Cambridge University Press, 2001.

[3] M. Falkenstein, J.Hohnsbein, J. Hoormann, and L. Blanke. Effects of error in choice reaction tasks on the ERP under focused and divided attention. In C. Brunia, A. Gaillard, and A. Kok, editors, *Psychophysiological Brain Research*, pages 192–195. Tilburg Univerity Press, Tilburg, 1990.

[4] P. Ferrez. *Error-related EEG potentials in brain-computer interfaces*. Phd thesis, Thèse Ecole polytechnique fédérale de Lausanne EPFL, no 3928, 2007.

[5] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An Interior-Point Method for Large-Scale `1-Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[6] Z. J. Koles, J. C. Lind, and A. C. K. Soong. Spatio-temporal decomposition of the EEG: a general approach to the isolation and localization of sources. *Electroencephalography and Clinical Neurophysiology*, 95(4):219 – 230, 1995.

[7] A. Mazaheri, I. L. Nieuwenhuis, H. van Dijk, and O. Jensen. Prestimulus alpha and mu activity predicts failure to inhibit motor responses. 30(6):1791–1800, 2009.

[8] W. H. R. Miltner, C. H. Braun, and M. G. H. Coles. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *J. Cognitive Neuroscience*, 9(6):788–798, 1997.

[9] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40(3):636–666, 1998.

[10] P. A. Valdés-Sosa, M. Vega-Hernández, J. M. Sánchez-Bornot, E. Martínez-Montes, and M. A. Bobes. EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis. *Human Brain Mapping*, 30(6):1898 – 1910, 2009.

[11] M. van Gerven and O. Jensen. Attention modulations of posterior alpha as a control signal for two-dimensional brain-computer interfaces. *Journal of Neuroscience Methods*, 179(1):78 – 84, 2009.

[12] H. T. van Schie, R. B. Mars, M. G. H. Coles, and H. Bekkering. Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, 7(5):549–554, 2004.

EEG analysis for implicit tagging of video data

Sander Koelstra
MultiMedia and Vision Group
Queen Mary, University of London, UK
Sander.Koelstra@elec.qmul.ac.uk

Christian Mühl
Human Media Interaction
University of Twente, NL
C.Muehl@utwente.nl

Ioannis Patras
MultiMedia and Vision Group
Queen Mary, University of London, UK
I.Patras@elec.qmul.ac.uk

Abstract

In this work, we aim to find neuro-physiological indicators to validate tags attached to video content. Subjects are shown a video and a tag and we aim to determine whether the shown tag was congruent with the presented video by detecting the occurrence of an N400 event-related potential. Tag validation could be used in conjunction with a vision-based recognition system as a feedback mechanism to improve the classification accuracy for multimedia indexing and retrieval. An advantage of using the EEG modality for tag validation is that it is a way of performing implicit tagging. This means it can be performed while the user is passively watching the video. Independent Component Analysis and repeated measures ANOVA are used for analysis. Our experimental results show a clear occurrence of the N400 and a significant difference in N400 activation between matching and non-matching tags.

1. Introduction

Given the enormous amount of unannotated multimedia data available nowadays, the need for automatic categorisation and labelling of video material to enable efficient indexing and retrieval is evident. So far, the predominant method used for tagging video data is by manual annotation. This is a slow, labour intensive process that cannot keep up with the amount of newly generated multimedia data. Lately, research has focused on finding ways to automate the annotation of this data. The use of EEG in this process is interesting mainly because it offers the possibility of passive, implicit tagging. This means that tags can be generated by analysing the EEG data as subjects consume multimedia data, without active involvement or conscious effort on their part. While at the moment the recording of EEG measurements is still a quite cumbersome process, recent improvements in the development of dry electrodes may simplify the use of this modality and make it usable outside of the laboratory environment.

The use of EEG in annotating multimedia data is a very new research direction and so far only a few works have investigated this area. In [6], an oddball paradigm is used in which images of a forest environment were shown to subjects for 100 ms each. The goal was to detect a small subset of target images that contained pedestrians. The target images elicit a P300 event-related potential which was then classified using Fisher linear discriminant analysis. Another test was run without the EEG modality, where subjects pressed a button upon seeing the target images. The results showed no significant differences in target image detection accuracy between the use of the EEG modality and the use of buttons. In [8], categories of images are classified based on EEG measurements recorded as the images were presented. The used categories were faces, animals and inanimate objects. This was based on the notion that the human visual system responds very differently to these categories of images. The authors propose a vision-based algorithm that uses pyramid match kernels to initially classify the images. The EEG data is then combined with the vision-based features using a kernel-alignment method. The combination of the two modalities outperforms the individual methods. In [3] the RAPID system is proposed. The authors use ERP analysis in combination with eye tracking to assist intelligence analysts in rapidly reviewing and categorizing satellite imagery. The analyst is assigned a target category to look for in the images. When subjects see an image in the target category, an ERP occurs in the EEG data which is then classified. Eye tracking is used to determine points of interest within the images.

All of these works are based on image annotation where as we attempt validation of tags related to video data. Also, in contrast to these earlier works, we perform tag validation rather than trying to assign tags directly. We show that there are significant differences between the cases of matching and non-matching tag presentations. This approach can be used in combination with a vision-based indexing and retrieval system in order to validate and re-rank its output, or for validating tags added manually by users. Such a tag validation system could be especially helpful in cases were the content to be tagged and the label categories are too com-

978-1-4244-4799-2/09/$25.00

plex (and only obvious from the incorporation of a wider context) to be classified by machine learning from the media directly. In that case the human (neural) responses can be used to indirectly classify the material. Many actions, such as for instance greeting a person, can vary greatly (e.g. waving, handshaking, hugging etc.) and be very difficult to detect via machine learning techniques. However, a human observer will have no difficulty in recognising these actions.

Another possible application is be the automatic recognition of social or affective content. In [9] an N400 response was observed for labels presented after musical excerpts. These words were very loosely attributed to the music in terms of associated objects (e.g. birds, needles), musical features, and moods. While these sub-categories were not analysed and reported separately, it is conceivable that the label information can entail categories of emotional content. As emotions are subjective in nature, the N400 approach to tag validation introduced here could in principle assess the subjective response to media content, thereby crossing a threshold insurmountable by a direct media analysis.

2. Methodology

We propose an approach to implicit tag validation through the use of EEG signals. In this approach, a subject is shown a video followed by a tag, and from the EEG signals recorded during tag display, we aim to discern whether the tag applies to the video content or not. Our hypothesis is that if the shown tag does not match the video content a 'mismatch negativity' will occur in the form of an N400 event-related potential (ERP). It has been shown that in cases of two semantically mismatching categories an N400 event-related potential occurs at around 400 ms after the second stimulus is presented (or better: after the mismatch becomes obvious to the viewer). This N400 has been observed even when the stimuli originate from different modalities (e.g. audio and text or images and text) [14, 12, 9, 1]. We aim to show here that the mismatch negativity can also be observed when we combine the modalities of video and text by priming the subject by the display of video content, followed by the display of a semantically mismatched tag. To the best of our knowledge this is the first work combining the video and text (tag) modalities.

We collected a large dataset with 17 subjects, each recorded for 98 trials. We use independent component analysis to remove eye blinks and other artefacts in the data and then determine whether the signals for the two cases (matching and non-matching tags) are significantly different using a repeated measures ANOVA. We found that there are indeed significant differences in the signal between the two cases in certain areas of the brain. We will now describe each step of our analysis in detail.

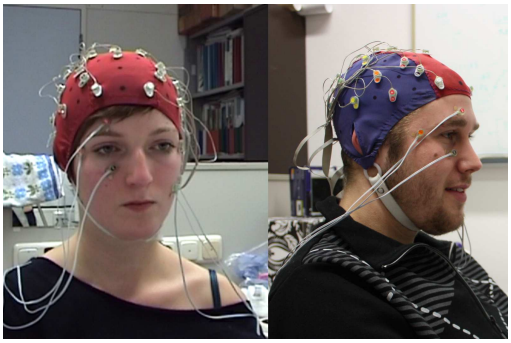


Figure 1. Subjects performing the experiment.

2.1. Experiment Setup

EEG was recorded using a Biosemi ActiveTwo system (www.biosemi.com) on a dedicated recording PC (P4, 3.2 GHz) using the BioSemi Actiview recording software. Stimuli were presented on a dedicated stimulus PC (P4, 3.2GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli the Presentation software by Neurobehavioral systems (www.neurobs.com) was used. Subjects were seated in a comfortable chair, approximately 70 cm from the presentation monitor (a 20 inch Samsung Syncmaster 203B). In order to minimise eye movements, the video stimuli were all shown width a width of 640 pixels, filling approximately a quarter of the screen. Each subject signed an informed consent form and filled in a short questionnaire. They were then instructed to try to restrict any movement to the periods between trials to minimize movement artefacts in the EEG signal. Subjects were told they would be shown videos followed by tags, but were not given any further specific instructions as to the nature of the experiment. 32 active AgCl electrodes were used (placed according to the international 10-20 system) and the data was recorded at 512 Hz. Fig. 1 shows two subjects as they perform the experiment.

17 Subjects were each recorded for 98 consecutive trials. 12 subjects were male, 5 female. Ages ranged from 19 to 31, with a mean age of 25. All but two subjects were right-handed and all but three subjects viewed the tags in their native language. Each trial consisted of the following steps:

1. A fixation cross is displayed for 1000 ms (to minimise eye movements).
2. The video is displayed (ranging in duration from 6-10 seconds).
3. A fixation cross is displayed for 500 ms.
4. The tag is displayed for 1000 ms.
5. A fixation cross is displayed for 4000 ms before the start of the next trial.

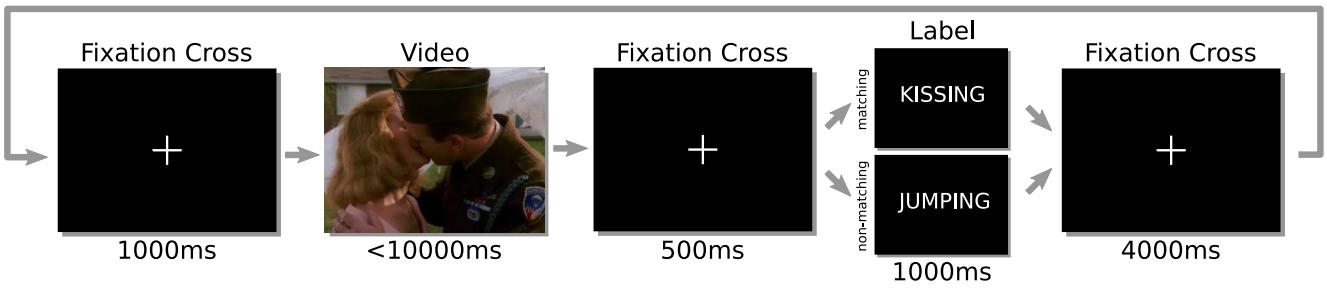


Figure 2. Order and timing of the experiment.

The stimuli were presented in 3 blocks of 32-33 trials. In between the blocks, subjects were given breaks and could move freely, reseat themselves or have a drink of water in order to avoid any muscle straining or fatigue. Fig. 2 illustrates the order and timing of the experiment.

49 Videos from seven different categories were used as stimuli, with 7 videos in each of the 7 categories. Each video has a duration of ten seconds or less and was shown twice, once followed by a matching tag and once followed by an incorrect tag. Table 1 gives an overview of the different video categories and their sources. The categories were chosen according to two criteria. Firstly, the categories should encompass events which do not vary too much in appearance within one category (to facilitate an eventual vision-based analysis). Secondly, we selected categories with human faces, animals and inanimate objects, following [8], who indicate that these categories can be separated reasonably well by analysing the EEG signals from subjects watching the videos.

2.2. Analysis

As a preprocessing step, the data was referenced the common average (CAR). Also, the data was bandpass-filtered between 0.5 and 40Hz to remove DC drifts and suppress the 60Hz power line interference. We extracted epochs for further analysis ranging from 500 ms before tag display to 1000 ms after. To remove interference caused by eye blinking and other artefacts, we perform spatial filtering using Independent Component Analysis (ICA). ICA has been used before in EEG data analysis with good results (e.g. [7]). Components containing only noise were manually selected and removed from the data. Fig. 3(a) is an example of a component that is strongly correlated with eye blinks. This is evident because the activation occurs in isolated periods (blinks) that are not correlated across trials. Also, the component is mostly active in the frontal electrodes. Such components are removed. Fig. 3(b) shows an example component correlated with the N100 and P200 ERP. The activation is concentrated in the occipital lobe (which is concerned with vision tasks), the component shows a resemblance to a typical ERP curve and there is a strong correlation between trials.

After removing the components that are due to blinks and other artefacts, we perform a repeated measures ANOVA to determine whether significant differences occur in the recorded EEG signal between the cases of matching and non-matching tags. For this purpose, we only consider the period of 300-500 ms after tag display, during which the strongest N400 response can be expected.

3. Results

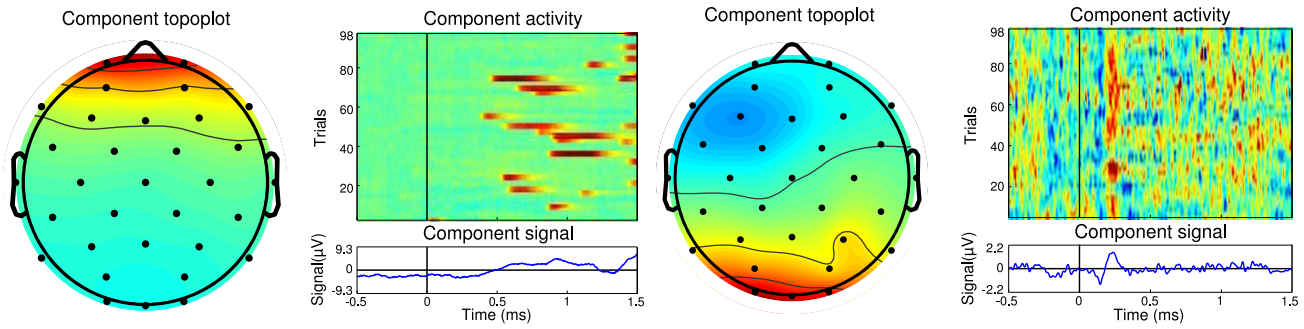
Table 2 shows the results of performing the repeated measures ANOVA. Results that have a p-value lower than 0.01 are deemed significant. Electrodes that show significant differences ($p \leq 0.01$) between the cases of matching tags and non-matching tags are highlighted. The fourth column shows the mean signal difference between the cases in μV (the mean signal in the case of matching tags minus the mean signal for the case of non-matching tags). Electrodes showing a significantly higher/lower negativity for non-matching tags are shaded light red and darker blue respectively.

Fig. 4 shows the location of observed differences in signal values. We can see that the differences are spatially mainly localised in two regions. The main region is located around the occipital and parietal lobe (covering electrodes CP1, Pz, PO3, CP2, C4 and Cz), where a more negative voltage deflection occurs when displaying non-matching tags than when displaying matching tags. The occipital lobe is concerned primarily with vision tasks and the parietal lobe is, among other things, concerned with the location of visual attention [10, 4]. The other region showing a significant difference in signal values is located in the left temporal lobe around electrodes AF3, FC5, T7 and F7. One of the functions of the left temporal lobe is the recognition of words, possibly explaining the activation there. In this case, the observed voltage is less negative for the case of non-matching tags than for the case of matching tags.

Fig. 5 depicts the grand average waveforms for the 9 electrodes exhibiting the most significant differences be-

Category/Label	Source
Airplane take off	Plane spotter homevideos (http://www.flightlevel350.com/)
People kissing	Hollywood movies dataset [11]
People getting out of cars	Hollywood movies dataset [11]
Mice drinking water	Mouse behaviour dataset [5]
Cats opening doors	Pet homevideos (http://www.youtube.com)
Jawdrop (posed facial expression)	MMI facial expression database [13]
Laughing people (spontaneous facial expression)	AMI meeting corpus [2]

Table 1. The different video event categories used in the experiment and their sources.



(a) An independent component that is strongly correlated with blinks. The component activity is concentrated in the frontal area and there is no correlation between trials.

(b) An independent component that is correlated with ERPs in the occipital cortex related with early visual processes. We can primarily see the activation here of the N100 and P200 ERP.

Figure 3. Visualisation of two independent components. In each of the subfigures: On the left is a topoplot of the component activation. In the top right the component activation is shown for 98 trials of one subject. In the lower right the average component signal is displayed.

tween the two cases. The first four plotted electrodes show less negativity for non-matching tags than for matching tags. The remaining electrodes show the opposite behaviour and display a higher negativity for the case of non-matching tags than for matching tags. Clear examples of the N400 ERP can be observed. The differences are most clear in the 300-500 ms period after tag display.

From these results it is clear that the N400 occurs when subjects are shown a combination of stimuli from the modalities of video and text (in the form of a tag). Furthermore, significant differences are present in a considerable number of electrodes between the cases of non-matching and matching tags. However, the effect size ($\leq 1\mu V$) is smaller than that found in other studies (e.g. [12, 1]). This can be due to the semantic categories, the stimulus material, or other parameters of the experiment used here.

4. Conclusions

In this work, we have collected and analysed a dataset to investigate the use of EEG for passive, implicit tag validation. Data was collected for 17 subjects and each subject was shown 98 videos, 49 followed by with matching tags and 49 followed by non-matching tags). Independent Component Analysis was used to remove noise (including eye blink artefacts) from the data. A repeated measures ANOVA showed significant differences in the EEG signal between the two cases of congruent and incongruent tags. This implies that the two cases can be successfully distinguished by analysis of the EEG signal. The next step in our research is to determine for single data trials whether the tag matches the video content. Successful single trial analysis would mean we can use this technique as a feedback mechanism in video analysis for indexing and retrieval. Other uses could include validating unreliable user-generated tags and possibly determining user reactions to the content (such as liking or disliking the content or other affective reactions).

In order to achieve a working tag validation system several parameters will have to be studied and optimized. Questions that need to be answered include: how long after a stimulus does a non-matching tags still elicit the ERP? What types of categories elicit the most robust mismatches? Does a subliminal presentation, not consciously perceived by the viewer, also elicit N400 responses? Can we also use a frequency analysis to judge how subjects implicitly judge the semantic meaning of the video?

In similar P300 experiments usually the EEG signal of several trials is averaged to increase the signal-to-noise ratio and increase the accuracy of ERP detection. This strategy could in principle also be used for the evaluation of label validity. However, it has to be ensured that multiple presented tags really are associated with the media content and

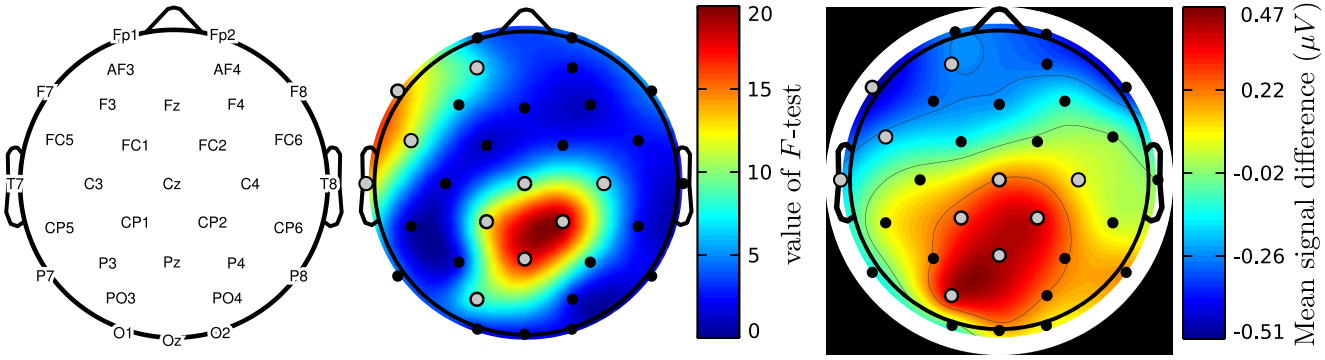


Figure 4. Left: Topoplot of Electrode locations, Middle: Topoplot of Significance of difference (F-test value), Right: Topoplot of the Grand-average differences between 300 and 500 ms for all 17 subjects. Electrodes with significant differences are highlighted in grey.

Electrode	$F(1, 16)$	p-value	MSD (μV)
CP2	19.98	0.000	0.776
Pz	17.59	0.000	0.819
CP1	11.74	0.001	0.535
Cz	08.15	0.004	0.480
PO3	07.32	0.007	0.616
C4	06.86	0.009	0.364
F7	15.25	0.000	-0.948
T7	14.15	0.000	-0.758
FC5	11.76	0.001	-0.640
AF3	07.84	0.005	-0.557
F3	06.50	0.011	-0.482
P4	06.30	0.012	0.478
P7	04.53	0.034	-0.443
F8	04.42	0.036	-0.455
Fp2	03.68	0.055	-0.417
Fp1	03.65	0.056	-0.429
FC2	02.29	0.130	0.260
Fz	01.69	0.194	-0.261
AF4	01.61	0.204	-0.250
FC6	01.57	0.210	0.203
CP6	01.33	0.249	0.236
P3	01.18	0.278	0.196
P8	01.09	0.297	0.209
O2	00.71	0.399	0.182
PO4	00.63	0.427	0.180
O1	00.63	0.428	-0.173
C3	00.43	0.514	0.094
F4	00.09	0.761	0.055
T8	00.08	0.783	0.053
Oz	00.06	0.808	0.056
CP5	00.03	0.870	-0.027
FC1	00.00	0.995	0.001

Table 2. ANOVA Results per electrode. MSD stands for Mean Signal Difference. Significant differences (p 0.01) are highlighted.

not with previously presented labels.

Using a single trial analysis, we hope to build a tag validation system that will achieve an efficiency close to that of manual tagging without active user involvement. However, given the low bitrate usually achieved by BCI systems, this task seems rather daunting. Also, mere tag validation does not compare to a complete manual tagging. Nevertheless, we envision a system that will be a useful addition to current tagging methods, especially given the absence of the requirement for active user involvement.

Acknowledgement

The research leading to these results has received funding from the Seventh Framework Programme under grant agreement no. FP7-216444 (PetaMedia).

References

[1] V. Bostanov and B. Kotchoubey. The t-CWT: A new ERP detection and quantification method based on the continuous wavelet transform and Students t-statistics. *Clinical Neurophysiology*, 117(12):2627–2644, 2006.

[2] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

[3] A. Cowell, K. Hale, C. Berka, S. Fuchs, A. Baskin, D. Jones, G. Davis, R. Johnson, R. Patch, and E. Marshall. Brainwave-Based Imagery Analysis. *Digital Human Modeling: Trends in Human Algorithms*, pages 17–27, 2008.

[4] A. Cummings, R. Čeponienė, A. Koyama, A. Saygin, J. Townsend, and F. Dick. Auditory semantic networks for words and natural sounds. *Brain research*, 1115(1):92–107, 2006.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] A. Gerson, L. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Trans. Neural Systems and Rehabilitation Engineering*, 14(2):174–179, 2006.

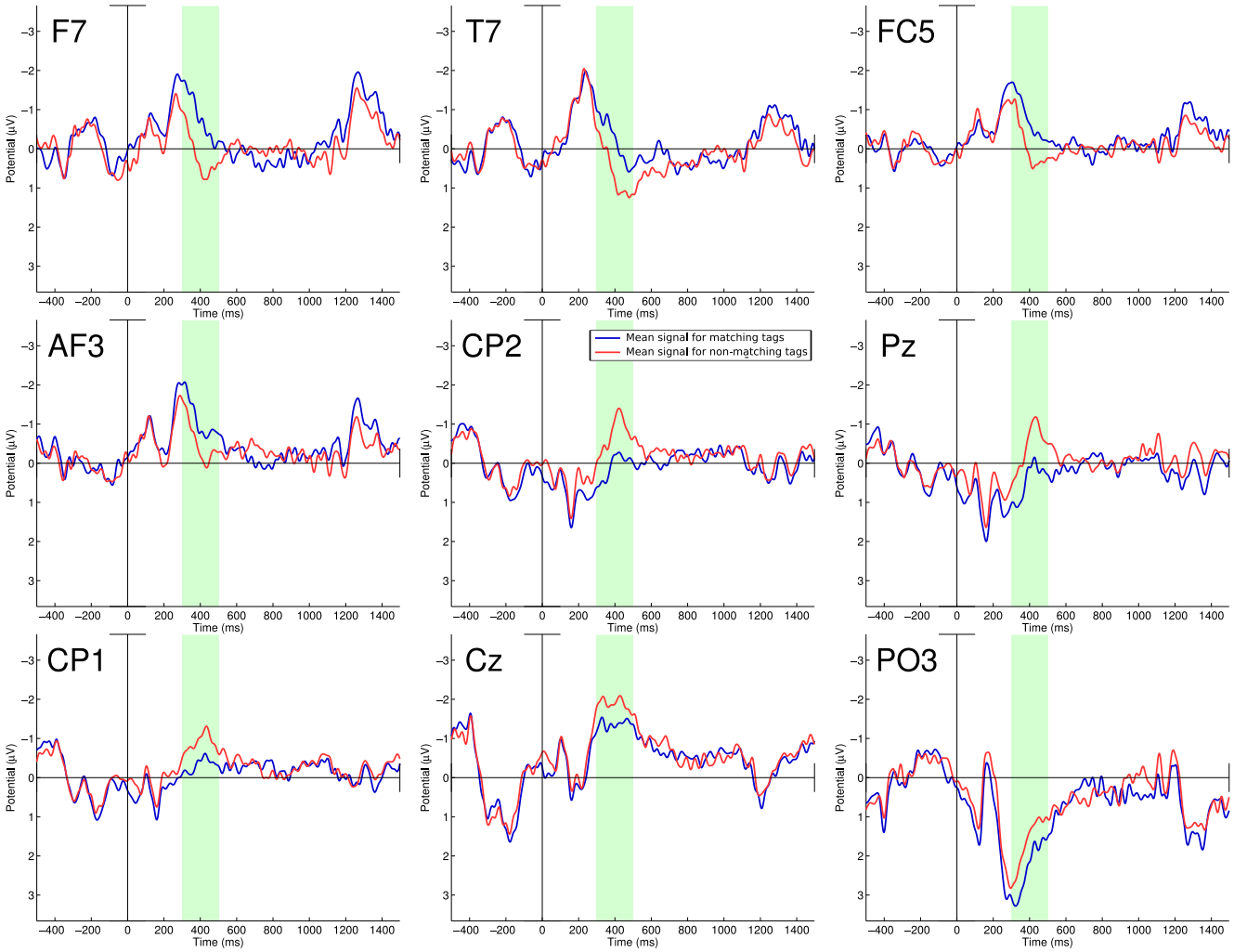


Figure 5. Grand average waveforms for the period 500 ms before to 1500 ms after tag presentation for the 9 electrodes with the most significant differences. The signal is averaged over all trials and subjects. The red line shows the average signal during presentation of matching tags and the blue line shows the average signal for non-matching tags. Differences in signal values between the two categories can be observed in each plot around the 400 ms mark. The light-green shaded area is the window used for ANOVA analysis. A 30 Hz low-pass filter was used in these plots, for display purposes only. Note that for the y-axis, negative is up.

[7] B. Kamousi, Z. Liu, and B. He. Classification of motor imagery tasks for brain-computer interface applications by means of two equivalent dipoles analysis. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 13(2):166–171, 2005.

[8] A. Kapoor, P. Shenoy, and D. Tan. Combining Brain Computer Interfaces with Vision for Object Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. Friederici. Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7:302–307, 2004.

[10] M. Kutas and S. Hillyard. Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, 11(5):539–550, 1983.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[12] G. Orgs, K. Lange, J. Dombrowski, and M. Heil. Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology*, 65(2):162–166, 2007.

[13] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int. Conference on Multimedia and Expo*, pages 317–321, July 2005.

[14] C. van Petten and H. Rheinfelder. Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, 33(4):485–508, 1995.

Measuring Task Engagement as an Input to Physiological Computing

Stephen H Fairclough
Liverpool John Moores University
Henry Cotton Campus, Liverpool
s.fairclough@ljmu.ac.uk

Katie C Ewing
Liverpool John Moores University
Henry Cotton Campus, Liverpool
k.c.ewing@ljmu.ac.uk

Jenna Roberts
Liverpool John Moores University
Henry Cotton Campus, Liverpool
j.roberts@ljmu.ac.uk

Abstract

Task engagement is a psychological dimension that describes effortful commitment to task goals. This is a multidimensional concept that combines cognition, motivation and emotion. This dimension may be important for the development of physiological computing systems that use real-time psychophysiology to monitor user state, particularly those systems seeking to optimise performance (e.g. adaptive automation, games, automatic tutoring). Two laboratory-based experiments were conducted to investigate measures of task engagement, based on EEG, pupilometry and blood pressure. The first study exposed participants to increased levels of memory load whereas the second used performance feedback to either engage (success feedback) or disengage (failure feedback) participants. EEG variables, such as frontal theta and asymmetry, were sensitive to disengagement due to cognitive load (experiment 1) whilst changes in systolic blood pressure were sensitive to feedback of task success. Implications for the development of physiological computing systems are discussed.

Introduction

Physiological computing (PC) describes systems that capture psychophysiological changes in the user in order to inform real-time software adaptation [1, 2]. PC systems rely on psychophysiology to create a representation of the psychological state of the user in real-time, e.g. changes in cognitive activity, positive and negative emotions, high vs. low task motivation. The system consults this representation to select an appropriate category of adaptive response. For example, if the user is frustrated, changes in user state should prompt the presentation of help information; if a player is bored by a computer game, the representation of user state should trigger an increase of game difficulty [3, 4]. The purpose of this approach is to create real-time software adaptation that is both implicit and intuitive.

The PC paradigm encompasses several existing strands of research/applications, from the control of adaptive automation [5, 6] to the use of psychophysiology to represent user emotion [7]. Unlike BCI applications [8], the PC approach is essentially passive (i.e. requiring no additional activity on the part of the user) and works mainly at the meta-level of the human-computer interaction (HCI) (i.e. ensuring that negative psychological states are minimised), i.e. whereas BCI represent an alternative form of input control [9].

The cycle of data collection and system response wherein psychophysiological change is transformed into adaptive control may be described as a biocybernetic loop [10]. This category of biocybernetic system control creates a symmetrical form of HCI where the availability of system information to the user is balanced by data about user state being at the disposal of the system [11]. Making a computer system privy to psychophysiological states has the potential to enable so-called 'smart' technology, i.e. systems that are characterised by increased autonomy and adaptive capability [12]. If technology develops in this direction, there is a subtle shift in the dynamics of HCI, from the master-slave dyad that characterises the way we currently use computers towards a collaborative, symbiotic relationship that requires computer technology to extend awareness of the user in real-time [13, 14].

One fundamental question surrounding the development of PC systems concerns how best to operationalise and represent the user state. There are several aspects to the question that should be considered during the initial stage of system design. In the first instance, what kind or dimension of user state is the most important one for a particular application domain? For example, physiological computing systems designed to control automation in the aircraft or vehicle cockpit have traditionally been concerned with representing the cognitive capability of the operator, specifically the prevention of Hazardous States of Awareness (HSA) [15]. Systems that employ psychophysiological measures for affective computing application emphasise the monitoring of negative affective states, such as anxiety [16] and frustration [17]. Similarly, psychophysiological monitoring has been used to identify quasi-emotional states, such as enjoyment, for those investigating this approach in the context of computer games [18]. At the second stage of system design, the researcher must identify those psychophysiological measures that provide the best operationalisation of the required psychological dimension. This stage may involve perusal of background literature followed by a series of validation experiments in the laboratory or the field, see [2] for full description of these issues.

This paper is concerned with how to measure the psychological dimension of task engagement as the basis for the development of PC systems. Task engagement is defined as "effortful striving towards task

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

goals" [19]. This multidimensional concept incorporates at least three psychological dimensions: (1) the investment of mental effort to optimise cognitive performance, (2) motivation to successfully achieve task goals, and (3) affective changes associated with the likelihood of goal attainment. This dimension is important because engagement has a predictive relationship with human performance (i.e. greater engagement = superior performance) and wellbeing (i.e. disengagement from a task is associated with negative psychological states such as boredom or anxiety).

Previous research

Research into biocybernetic control of adaptive automation at NASA focused on the measurement of spontaneous electroencelographic (EEG) activity in order to capture task engagement, i.e. an EEG index ratio measure where the ratio of mean power in the high-frequency beta bandwidth (13-40Hz) is divided by total power in lower-frequency alpha (8-12Hz) and theta (3-7Hz) components ($\beta/(\alpha+\Theta)$) [10]. This prototypical system enabled automation of a laboratory-based task (the Multi-Attribute Task Battery - MATB) provided that the operator was deemed to exhibit high task engagement; if EEG measures of task engagement went into a decline whilst automation was activated, the system switched the user into a manual control mode, i.e. to re-engage with the task and prevent automation-induced complacency. This programme of research is summarised in [20].

The measurement of task engagement using psychophysiology takes on a different complexion in the context of desktop-based systems. For example, detection of negative user states is particularly relevant for computing applications designed to aid learning [21]. Recent work on the detection of user frustration [17] demonstrated the utility of the multimodal approach that combined multiple measures to predict subjective feelings of frustration. These authors measured skin conductance in combination with posture analysis, detection of head gestures (head shakes and nods), facial expression (smiling) and haptic monitoring. These measures were used to predict self-reported episodes of frustration, which was accurately detected in 79% of all cases (chance level = 58%). This experiment demonstrated how covert psychophysiology may be combined with overt behavioural signals in order to define the psychological dimension of interest. Related work on affective computing has also combined different psychological dimensions to yield a suitable representation of user state. For example, Burleson and Picard [22] described a state of "stuck" that may occur during the learning process to the detriment of user motivation. The definition of this state combines negative affect (e.g. anxiety) with cognitive characteristics (e.g. inability to focus, mental fatigue).

Measuring task engagement via psychophysiology

Task engagement can be defined with respect to cognitive activity (mental effort), motivational orientation (approach vs. avoidance) and affective changes (positive vs. negative valence).

Mental effort is conceptualised as energy mobilisation in the service of cognitive tasks or goals. At the cerebral level, the electrical activity of the brain may be quantified via the EEG to study how different states of brain activation represent the level of mental effort investment. The topography of EEG activation may provide important information about the specificity and distribution of activation over the cortex. A series of experiments demonstrated that augmentation of theta activity (4-7Hz) from central frontal sites and suppression of alpha activity from occipital areas were both associated with increased mental effort in response to working memory load (i.e. number of items to be retained in memory) [23, 24].

The pupillary response has a long association with the measurement of mental effort in response to cognitive variables [25, 26]. There is evidence that pupil dilation is greater during the processing of a complex cognitive operation relative to a simple one. The main problem with pupilometry is interference from light adaptation, i.e. for those environments where the level of lighting is not carefully controlled. The Index of Cognitive Activity [27]represents an attempt to quantify small discontinuities in pupil size that are related to cognitive activity. The ICA is derived in a selective manner that minimises the influence of lighting levels.

There is an obvious link between task engagement and the motivation to successfully achieve a given outcome. Motivational intensity theory [28, 29] proposes that goal commitment (i.e. the willingness to invest effort into the task) is a function of perceived: (i) task difficulty, (ii) ability, and (iii) likelihood that successful performance on the task will achieve a desired motive (e.g. monetary incentives, prowess, 'feeling good'). Therefore, if the individual assesses themselves to have the requisite level of skill to achieve success, then effort is invested into performance. Research into motivational intensity theory has used indicators of sympathetic nervous system (usually systolic blood pressure) to describe the "tipping point" where increased difficulty/reduced perception of ability/reduced perception that the task is worthwhile forces participants to switch from effortful striving for goal success to disengagement and a significant reduction of mental effort [30, 31].

Related research has linked changes in frontal EEG asymmetry to the self-regulation of affect and motivational orientation. In broad terms, the experience of positive emotions is associated with high levels of relative left frontal activity, whereas negative emotions is related to increased relative right frontal activity [32, 33]. There is also evidence that increased left frontal

activation is correlated with motivational approach whilst right frontal activation is linked with a motivation disposition in the direction of avoidance. Research into the influence of reward on frontal asymmetry supports this connection [34-36], and higher levels of left frontal activation have been associated with trait measures of behavioural activation [37-39]. The relationship between motivational direction and affective valence encapsulated by frontal EEG asymmetry is implicit within a performance setting.

Task engagement is a multidimensional description of user state [40] that incorporates psychophysiological measures of cognition, motivation and affect. The relationship between physiology and psychology may be described as many-to-one [41] as multiple indicators from EEG, pupilometry and cardiovascular activity are deployed in concert to represent this dimension of task engagement. The purpose of this paper is to describe two laboratory experiments, both dedicated to the measurement of task engagement using different types of manipulation. In experiment one, participants are exposed to five levels of task demand using a working memory task. The aim of this experiment is to mentally overload the participants so he or she decides to withdraw effort from the task because it is deemed to be too difficult to achieve. The second experiment manipulated task engagement by providing participants with false feedback about the quality of their performance. One group was informed that performance was successively improving over time whereas the second group of participants received feedback of progressive performance decline. In the case of the second experiment, task engagement is influenced by manipulating participants' perception of their own ability.

Experiment 1: Mental Overload

Description of Study

21 participants (11 male) took part in the research, however data from 3 participants was excluded due to EEG artefacts and incorrect task completion. Participants were aged between 19 and 39 years of age. Cognitive effort was elicited with a verbal working memory task known as the n-back task. The task requires participants to indicate if the currently presented stimulus matches one shown on an earlier occasion. Solid black letters (against a white background) were presented to participants on colour monitor at a distance of 80cm. The task consisted of 6 levels of difficulty, with level 1 being the easiest and level 6 the most difficult. For each stimulus presentation participants needed to indicate if the letter matched the previous letter (level 1), the letter 2-previous (level 2), the letter 3-previous (level 3), the letter 4-previous (level 4), the letter 5-previous (level 5) and the letter 6-previous (level 6). Responses were given with a keyboard press of 1 for match and 2 for non-match, using the right index and middle fingers. Participants attended a training session of approximately 4.5 hours on the day before the experiment.

EEG activity was recorded monopolarly from 32 Ag-AgCl pin-type active electrodes mounted in a BioSemi stretch-lycra headcap. Electrodes were positioned according to the international 10-20 system and EEG activity recorded from the following sites: frontal pole (FP1, FP2), Anterior-frontal (AF3, AF4), frontal (F3, Fz, F4), fronto-central (FC5, FC1, FC2, FC6), central (C3, Cz, C4), temporal (T7, T8), parieto-central (CP5, CP1, CP2, CP6), parietal (P7, P3, Pz, P4, P8), occipito-parietal (PO3, PO4) and occipital (O1, Oz, O2). Electrodes were also placed at earlobe sites (A1, A2) allowing electrodes to be referenced off line to a linked ears reference. EEG was recorded continuously throughout a 4 minute baseline prior to the task and continuously throughout the task.

Systolic blood pressure measurements were taken using a Dinamap Vital Signs monitor (PRO100) using a cuff that was worn on the upper arm. Readings of systolic, and diastolic blood pressure along with heart rate and mean arterial pressure were obtained. A baseline reading was taken during a 4 minute period prior to task completion at 180s after the start of this period. Readings were then taken for each experimental trial 60s after onset giving 2 readings for each task level.

Pupil diameter measurements were recorded continuously at a sample rate of 60Hz with two remote infrared video cameras (Seeing Machines Ltd, Canberra, Australia). The cameras used binocular tracking and were mounted on a metal frame 80-90cm in front of the participant, placed beneath the stimulus display monitor. Pupil size resolution was possible at 0.00001mm. Data was recorded using FaceLAB 4.6 software. Illumination from the stimulus display and room lighting (8 x 36W ceiling mounted fluorescent tubes) was maintained within the range of 355-380Lux at the seated position of the participant to avoid a confound with the pupillary light reflex. Pupil diameter was measured throughout a 2min baseline prior to task completion during which participants were required to maintain their gaze at a fixation point (green dot) at the screen centre. Measurements were then made continuously throughout each trial. Participants were asked to keep still and maintain fixation at the centre of the screen minimizing possibility of head movement artifacts in the signal.

All EEG analysis was performed using BESA software (MEGIS software GmbH, Gräfelfing, Germany). First a 50Hz notch filter was applied to the raw data along with a 0.05Hz high pass and 60Hz low pass filter. A linked ears montage was applied. Data was visually inspected for artefacts from external electromagnetic sources which were excluded. Data underwent automatic correction for blink artefacts, horizontal and vertical saccades based on detection through predefined topographies. Average power spectra were then computed for each experimental condition. Power spectra in μV^2 were Log transformed

(natural log) to normalise the distribution. Frontal asymmetry values were obtained for all 7 experimental conditions using EEG power values from the following electrode sites: FP2, AF4, F8, F4, FC2, FC6, C4, T8,(right hemisphere sites) FP1, AF3, F7, F3, FC1, FC5, C3, T7. (left hemisphere sites).

Power estimates for frequencies lying within Individual Alpha Bands were then used in the following formula: Ln [right total alpha power] – Ln [left total alpha power] to generate an asymmetry index [42]. Positive values indicated greater relative right alpha power and greater relative left frontal activity, greater relative right frontal activity was indicated by negative values. Asymmetries were also calculated for homologous pairs of electrodes.

Data from the left and right eye of 14 participants (7 female) was pre-processed to remove erroneous measures of pupil diameter arising from blinks, partial blinks, electromagnetic noise and artefacts resulting from tracking failure and camera joggle. Readings of 0 or near 0 were eliminated from the data to exclude blinks, partial blinks and tracking failure, and readings differing by more than +/- 0.1mm from the previous observation were excluded to reduce the influence of noise. The data then underwent 1-D wavelet decomposition using the orthogonal wavelet 'db4' from the Daubechies family of wavelets. The decomposition was achieved by convolving the signal with a high and low pass filter followed by downsampling by a factor of 2. Decomposition was performed using 5 iterations on each signal. The procedure produced a set of detail coefficients which were subjected to a minimax (hard) threshold to reduce noise, in which noise was presumed to be Gaussian white noise. Detail coefficients were then subjected to a threshold of 0.05 and coefficients above this value interpreted as showing high frequency discontinuous increases in pupil diameter. Numbers of these discontinuities, which have been found to correlate with cognitive processing [27], were used to generate an index consisting of the average no of discontinuities per second for each condition.

Results

EEG data were analysed with respect to two primary variables: frontal theta activity from the central area (Fz) and frontal asymmetry data. Theta activity at Fz was calculated using the dominant frequency (i.e. as personalised to each individual). The average power at the dominant theta frequency was calculated and submitted to analysis via ANOVA. The results revealed a significant trend [$\underline{F}(6,12)=3.09$, $\underline{p}<0.05$]. Post-hoc testing revealed that theta activity was significantly lower at baseline, the one-back and the six-back task compared to all other conditions ($\underline{p}<0.05$).

Activity in the alpha bandwidth was also calculated with respect to the dominant frequency. Alpha power at the dominant frequency was calculated for all participants and converted via natural log prior to analysis. Asymmetry scores (left side minus right side) were calculated across three pairs of frontal sites on either side of the midline: AF3-AF4, F3-F4, FC3-FC4. Therefore, an increase of the asymmetry score is equated with greater activation of the left hemisphere. Each asymmetry score was analysed using an ANOVA model. There were no significant results for those asymmetry scores calculated with AF3-AF4 or F3-F4; however, the frontal-central sites (FC3-FC4) revealed a significant trend [$\underline{F}(6,12)=2.57$, $\underline{p}<0.05$]. Post-hoc testing revealed greater left-hemisphere activation (i.e. approach motivation) during all task conditions compared to baseline or the six-back condition ($\underline{p}<0.05$). In other words, both the baseline (resting) condition and the six-back task were associated with greater levels of right hemispheric activation, which is associated with avoidance motivation. This finding is illustrated in Figure 1.

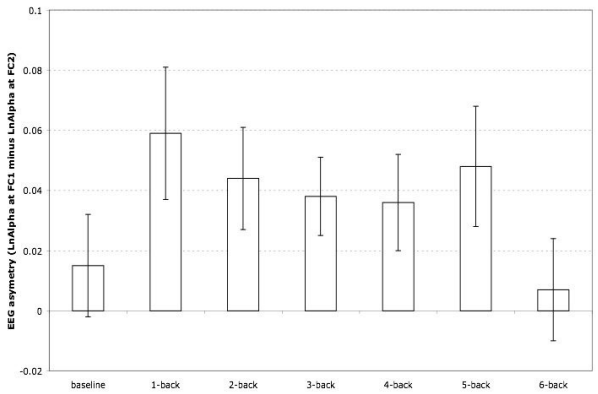


Figure 1. Frontal asymmetry scores (left side alpha power minus right side alpha power at dominant frequency) for FC3-FC4 across all six task demand conditions (N=18).

The measurement of systolic blood pressure has been associated with mental effort and task motivation. The analysis of this variable revealed a significant trend [$\underline{F}(6,12)=13.01$, $\underline{p}<0.01$]; however, post-hoc testing revealed only a significant difference between resting baseline and task conditions, i.e. the measure failed to distinguish between different levels of task demand.

An approximation of the Index of Cognitive Activity (ICA) was calculated for 14 participants based on changes in the pupil size. Specifically, the ICA captures short discontinuities in pupil size related to changes in mental workload. These data were subjected to ANOVA analysis, which revealed a significant difference due to experimental condition [$\underline{F}(6,8)=7.26$, $\underline{p}<0.05$]. Post-hoc testing revealed that the ICA was significantly lower than all working memory conditions, i.e. the ICA was not significantly sensitive to changes in working memory load (see Figure 2).

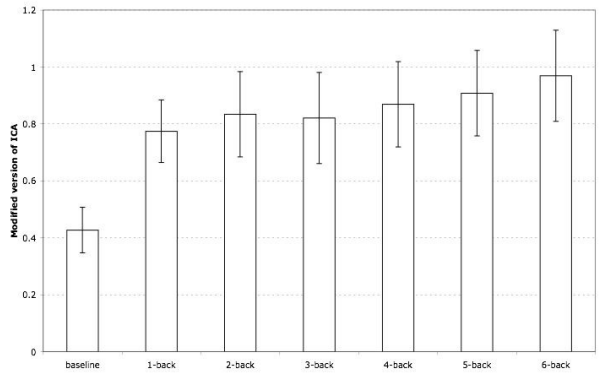


Figure 2. Mean score on modified Index of Cognitive Activity measure(N=14).

Experiment 2: Performance Feedback

Description of Study

34 participants (17 males and 17 females) formed 2 independent groups. A positive feedback group completed a working memory task and received pre arranged performance scores indicating a gradual improvement in performance over time. A negative feedback group completed the same task but received scores indicating a gradual decline in performance.

The memory task was computer based and was created using E-Studio software. It was developed from the 'n-back task' [24] . The version of the task used in this study was a 2 back task where participants continuously compared a currently presented stimulus to one seen 2 trials previously. Participants were presented with a 3x3 grid. On each trial, a green square appeared at one of the 9 grid locations for 1.75 seconds and was immediately followed by the next square. Participants were asked to respond on every trial by pressing 1 of 2 keyboard buttons to indicate that the location of the current square was either in the same location as the square seen 2 trials before (a match) or in a different location (a mis-match). The task was divided into 5 blocks, each of which contained 90 trials. Each block lasted just over 2.5 minutes and matches occurred on approximately 35% of trials.

In the experimental session of the study, participants were provided with false performance feedback as a percentage of overall accuracy at the end of each task block. Performance feedback was presented via a second computer placed adjacent to the memory task computer. Participants were misled to believe that performance data was being calculated in real-time by this second computer following each block of task activity. This illusion was achieved via a macro written in Microsoft Excel. The macro simulated a process of calculation and analysis and produced a chart to display performance accuracy. Each chart also included performance levels from any previous block/s which provided a visual representation of a gradual decline in performance for the negative feedback group and a gradual improvement in performance for the positive feedback group. Both groups received performance feedback of 60% after block 1 and both groups showed a cumulative decline or increase of 11% in total from block 1 to 5. For the negative feedback group, performance accuracy scores fell from 60% after block 1 to 56% after block 2, to 53% after block 3, to 52% after block 4 and finally reached 49% after block 5.

Blood pressure was recorded using a standard Dinamap with the pressure cuff placed over the brachial region of the participant's left arm. Initial screen and baseline readings were taken at the start of the experiment. Whilst participants worked on the memory task, 2 blood pressure readings were taken after approximately 20 seconds and 120 seconds from which an average was calculated.

EEG was recorded using active electrodes and sampled at 512Hz via a BioSemi system. Offline, EEG signals were corrected for ocular and physical artifacts and filtered using high and low band pass filters of 0.16Hz and 15Hz respectively. Artifact free epochs were then analysed via Fast Fourier Transform which yielded mean power in the alpha (8-12Hz) bandwidth. Alpha activity in the right hemisphere relative to homologous left hemisphere sites was calculated (ln [right] – ln [left]) to produce scores of alpha asymmetry for the following pairs of frontal sites: Fp2-1, Af4-3, F4-F3, FC2-FC1 and FC6-FC5. Theta activity was collected from frontal, central areas (Fz) as in the previous experiment.

Facial electromyographic activity (fEMG) was recorded to attain measures of muscle activity for the corrugator supercilii.

Results

EEG data: Two participants were excluded from this analysis due to technical problems with the data collection (one from each Feedback Group). The MANOVA analysis of EEG data revealed significant main effects for frontal asymmetry site, $F(4,26) = 5.70$, $p < .01$, and experimental condition, $F(1,29) = 4.05$, $p < .05$. The effect of experimental condition for EEG frontal asymmetry demonstrated that frontal asymmetry score (across all sites) was significant higher in the presence of performance feedback, i.e. higher level of activation in left hemispheric sites during feedback condition. There was no effect of feedback on levels of frontal theta activity.

Systolic Blood Pressure (SBP): The ANOVA model conducted on SBP data revealed a significant main effect for experimental condition, $F(1,30) = 4.82$, $p < .05$, i.e. SBP was significantly higher during Feedback [$\underline{M} = 115.78$] compared to the No Feedback condition

[M = 112.71]. The same model also revealed significant interactions between Feedback Group x Task Block, $F(4,27) = 3.55$, $p < .05$, and Feedback Group x Experimental Condition x Task Block, $F(4,27) = 3.20$, $p < .05$. For the positive feedback group, mean SBP was significantly higher at Task Block 5, $t(15) = 3.26$, during the Feedback condition compared to the No Feedback Condition (Figure 3).

Figure 3. Mean Systolic Blood Pressure (mm/Hg) for Positive Feedback Group compared across both experimental conditions (N=16).

Corrugator Activity: The corrugator data were subjected to ANOVA model with an eyes open baseline included as an additional cell in the Task Block factor. This analysis revealed no significant effects. The trend of the data was to increase in presence of Feedback and this trend was particularly prominent during Task Block 5.

Discussion & Conclusions

Explanation of findings

Two experiments were conducted to identify the sensitivity of psychophysiological variables to the manipulation of task engagement. In the first experiment, engagement was manipulated by systematically increasing task difficulty. It was anticipated that the high level of working memory load at the 5- and 6-back versions of the task would cause participants to disengage. However, there was evidence from subjective measures of workload (NASA-TLX) that a point of overload was not reached, i.e. mean TLX score at 6-back task = approx. 6.5 on a 10-point scale. A reduction of frontal theta and an increase of right hemispheric frontal activity (Figure 1) was observed at maximum task demand. These data indicated that our participants were reducing levels of mental effort and shifting motivational orientation towards avoidance. In other words, they were withdrawing from the task. The pupilometry data from the ICA did not yield a statistically significant trend, however, a trend was observed of increasing cognitive demand (Figure 2).

These findings beg a question about volitional vs. mandatory responses to task demand in the psychophysiological realm. The positive linear relationship between task demand and ICA illustrated in Figure 2 contradicts the quadratic pattern that characterised both frontal theta and EEG frontal asymmetry (Figure 1). We may speculate that the ICA represents a response to perceived task demand, regardless of engagement, whereas the quadratic trend describes a self-regulated process of energy mobilisation. With respect to the latter, the initial level of low task demand (e.g. 1-back task) failed to increase frontal theta, which increased rapidly for 2-, 3- and 4-back versions of the task, before falling during the highest levels of task demand. The trend for frontal asymmetry was slightly different (Figure 1); exposure to the task led to increased approach motivation (at the 1-back task), which declined as task difficulty increased (indicating avoidance motivation) with the exception of a marked increase at the 5-back version of the task.

The second experiment attempted to manipulate task engagement in two ways. First, it was anticipated that performance feedback inevitably increases task engagement as the quality of one's own performance is rendered more salient. By providing repeated exposure to both positive and negative feedback, we anticipated different patterns of mental effort investment; specifically, we expected positive feedback to reduce effort investment (as participants received the impression that performance was consistently improving).

It was hypothesised that the presentation of negative feedback would initially mobilise high levels of effort, leading to disengagement towards the latter periods of the task as prompted by repeated exposure to negative feedback. The first hypothesis was supported by the frontal asymmetry data; participants exhibited higher left frontal activation during the feedback condition (regardless of whether feedback was positive or negative). The only psychophysiological response to the direction of feedback was found with respect to systolic blood pressure. This variable is associated with sympathetic activation of the autonomic nervous system, i.e. increased activation. Whilst systolic blood pressure did not respond to different levels of task demand during the first experiment, this variable exhibited a broadly linear increase in response to feedback of positive performance (Figure 3). This pattern was unexpected but was interpreted in the following way; contrary to expectations, positive feedback increased participants' appraisal of their own capability, which motivated these individual to both aspire towards higher levels of performance and increase mental effort mobilisation. The absence of the opposite trend in the presence of negative feedback was puzzling; perhaps negative feedback had no impact on any psychophysiological indicators of effort because the task was quite abstract and there were no negative consequences of task failure

Implications for Physiological Computing

What conclusions can be drawn from these laboratory studies for the development of physiological computing (PC) systems? In the first instance, the pattern of EEG data from experiment one point to the feasibility of capturing task engagement as a volitional response to task demand. This may be particularly important for applications such as computer games, which emphasise both autonomy and different levels of task demand. It is proposed that theta activity at frontal-central sites and frontal asymmetry are investigated as real-time variables to be integrated into the biocybernetic loop. Both variables demonstrated a sufficient degree of sensitivity to justify follow-up work. Further research must also explore individualised algorithms using neural net approaches [43, 44].

It should be noted that both EEG variables failed to show any sensitivity to positive vs. negative performance feedback during the second experiment. Therefore, these EEG variables seemed to respond primarily to engagement in the context of cognitive load. On the other hand, systolic blood pressure, which demonstrated a sensitivity to performance feedback, failed to distinguish between different levels of cognitive load in the first experiment. This pattern of results demonstrates the multidimensional nature of task engagement - different categories of measures may exhibit sensitivity to specific aspects of the concept. In this case, EEG variables respond to disengagement due to cognitive load whilst changes in systolic blood pressure reacted to changes in goal-setting behaviour, i.e. a desire to achieve at a higher level.

The relationship between physiology and psychology may be described as 'many-to-one' in the case of task engagement [41]; data from several physiological sources are required to successfully capture this dimension. The data from both experiments demonstrate the sensitivity of certain variables to different levels of task load or performance feedback. But the crucial distinction for the development of PC systems is the discrimination between rising engagement, sustained engagement and sustained disengagement. Systems that are designed to adaptively respond to changes in engagement need to assess: (1) how to facilitate rising levels of task engagement, and (2) how to counteract periods where the user may become disengaged from the task. With respect to our data, systolic blood pressure would appear to be a candidate for (1) whereas the EEG variables were sensitive to (2).

From the perspective of system design, it is not simply a question of selecting the correct variable to represent engagement, there is also the issue of sensitivity to the specific aspect of task engagement that is central to the application. For designers of adaptive automation applications, it is important to protect safety-critical performance; therefore, the ability to detect and predict task disengagement is a top priority. If the PC approach is applied to an automatic tutoring system, detection of sustained or rising engagement becomes just as important because learning software should be designed to engross and inspire users, and to sustain these positive states via real-time adaptation.

It is important for designers to have a clear idea about the level of discrimination that the system must achieve in order to provide appropriate levels of adaptation. For some systems, detecting two categories of engagement will suffice (high vs. low engagement); other systems may require more fine-tuned levels of discrimination (high vs. high/med vs. med vs. med/low vs. low). As the number of possible categories increases, the quantitative distance between each category declines, which will lead to higher false positives or misses, so the designer must consider this trade-off to optimise the performance of the system as a whole. Much depends on the adaptive capability of the system under development, PC systems that are capable of only one kind of adaptation (e.g. present help vs. no help presentation) only require a two-category classification. Systems with several levels of adaptive capability (e.g. present four different categories of help information) will require a psychophysiological algorithm that can discriminate four levels of task engagement [2].

From the perspective of building PC systems, it is obvious that psychophysiological variables offer significant advantages for representing user states. These measures are covert, passive and highly sensitive, but this level of sensitivity is double-edged. Psychophysiological variables are sensitive to a wide range of possible influences from physical artifacts (moving the body) to environmental factors (room temperature) to diurnal influences (time of day), the effects of caffeine and food, exercise, personality, mood etc. If system designers wish to harness the sensitivity of psychophysiology, this double-edged property must be appreciated. One could resolve the problem by monitoring confounding variables in order to model and isolate their influence on the psycho-physiological inference that is central to PC systems. Alternatively, designers could seek context via another route by considering psychophysiological changes in the same data space as other categories of variable, i.e. a multimodal approach [45]. This approach would combine psychophysiological changes with behavioural markers, such as posture [46] and facial expression. Psychophysiological changes could also be assessed in relation to measures of task performance [47]. One could combine markers from several categories (psychophysiological, behavioural, performance) in order to discern the level of task engagement via a process of triangulation. The danger with this approach is how to handle divergence/disagreement between the different categories of data.

To conclude, task engagement is an important psychological dimension for the development of physiological computing systems. It is also a complex dimension incorporating aspects of cognition with self-regulatory activities such as goal-setting and motivation.

Laboratory experiments have been described to identify candidate variables such as EEG frontal asymmetry and systolic blood pressure. The next step is to evaluate these variables in the context of a computerised task in the field.

References

1. Allanson, J. and S.H. Fairclough, A research agenda for physiological computing. Interacting With Computers, 2004. 16: p. 857-878.
2. Fairclough, S.H., Fundamentals of Physiological Computing. Interacting With Computers, 2009. 21: p. 133-145.
3. Gilleade, K.M., A. Dix, and J. Allanson. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. in Proceedings of DiGRA 2005. 2005.
4. Fairclough, S.H. Psychophysiological inference and physiological computer games. in ACE Workshop - Brainplay'07: Brain-Computer Interfaces and Games. 2007.
5. Freeman, F.G., et al., Evaluation of a psychophysiologically controlled adaptive automation system, using performance on a tracking system. Applied Psychophysiology and Biofeedback, 2000. 25(2): p. 103-115.
6. Wilson, G.F. and C.A. Russell, Operator functional state classification using multiple psychophysiological features in an air traffic control task. Human Factors, 2003. 45(3): p. 381-389.
7. Picard, R.W., E. Vyzas, and J. Healey, Towards machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. 23(10): p. 1175-1191.
8. Allison, B.Z., E.W. Wolpaw, and J.R. Wolpaw, Brain-computer interface systems: progress and prospects. Expert Review of Medical Devices, 2007. 4(4): p. 463-474.
9. Fairclough, S.H., BCI and Physiological Computing: Similarities, Differences and Intuitive Control, in Workshop on BCI and Computer Games: CHI'08. 2008: Florence.
10. Pope, A.T., E.H. Bogart, and D.S. Bartolome, Biocybernetic system evaluates indices of operator engagement in automated task. Biological Psychology, 1995. 40: p. 187-195.
11. Hettinger, L.J., et al., Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. Theoretical Issues in Ergonomic Science, 2003. 4(1-2): p. 220-237.
12. Norman, D.A., The Design of Future Things. 2007, New York: Basic Books.
13. Klein, G., et al., Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intelligent Systems, 2004. 19(6): p. 91-95.
14. Pantic, M., et al., Human computing and machine understanding of human behaviour: a survey, in Artificial Intelligence for Human Computing, T. Huang, et al., Editors. 2007, Springer. p. 47-71.
15. Prinzel, L.J., Research on Hazardous States of Awareness and Physiological Factors in Aerospace Operations. 2002, NASA: Hampton, Virginia.
16. Rani, P., et al., Anxiety detecting robotic system - towards implicit human-robot collaboration. Robotica, 2004. 22: p. 85-95.
17. Kapoor, A., W. Burleson, and R.W. Picard, Automatic prediction of frustration. International Journal of Human-Computer Studies, 2007. 65: p. 724-736.
18. Mandryk, R.L. and M.S. Atkins, A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies, 2007. 65: p. 329-347.
19. Matthews, G., et al., Fundamental dimensions of subjective state in performance settings: Task engagement, distress and worry. Emotion, 2002. 2(4): p. 315-340.
20. Scerbo, M.W., F.G. Freeman, and P.J. Mikulka, A brain-based system for adaptive automation. Theoretical Issues in Ergonomic Science, 2003. 4(1-2): p. 200-219.
21. Picard, R.W., et al., Affective learning - a manifesto. BT Technology Journal, 2004. 22(4): p. 253-269.
22. Burleson, W. and R.W. Picard. Affective agents: sustaining motivation to learn through failure and a state of "stuck". in Workshop on Social and Emotional Intelligence in Learning Environments. 2004.
23. Gevins, A. and M.E. Smith, Neurophysiological measures of cognitive workload during human-computer interaction. Theoretical Issues in Ergonomic Science, 2003. 4(1-2): p. 113-121.
24. Gevins, A., et al., Monitoring working memory load during computer-based tasks with EEG pattern recognition models. Human Factors, 1998. 40(1): p. 79-91.
25. Beatty, J., Task-evoked pupillary responses, processing load and the structure of processing resources. Psychological Bulletin, 1982. 91(2): p. 276-292.
26. Beatty, J., Phasic not tonic pupillary responses vary with auditory vigilance performance. Psychophysiology, 1982. 19(2): p. 167-172.
27. Marshall, S.P., C.L. Davis, and S.R. Knust, The index of cognitive activity: estimating cognitive effort from pupil dilation. 2004, EyeTracking Inc.: San Diego, CA.
28. Wright, R.A. and J.D. Dill, Blood pressure responses and incentive appraisals as a function of perceived ability and objective task demand. Psychophysiology, 1993. 30: p. 152-160.
29. Wright, R.A. and A. Dismukes, Cardiovascular effects of experimentally induced efficacy (ability) appraisals at low and high levels of avoidant task demand. Psychophysiology, 1995. 32: p. 172-176.
30. Richter, M. and G.H.E. Gendolla, Incentive value, unclear task difficulty, and cardiovascular reactivity in active coping. International Journal of Psychophysiology, 2007. 63(3): p. 294-301.
31. Richter, P. and G.H.E. Gendolla, Incentive effects on cardiovascular reactivity in active coping with unclear task difficulty. International Journal of Psychophysiology, 2006. 61: p. 216-225.
32. Davidson, R.J., Anterior electrophysiological asymmetries, emotion and depression: conceptual and methodological conundrums. Psychophysiology, 1998. 35: p. 607-614.

33. Davidson, R.J., What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. Biological Psychology, 2004. 67: p. 219-233.
34. Pizzagelli, D., et al., Frontal brain asymmetry and reward responsiveness. Psychological Science, 2005. 16(10): p. 805-813.
35. Miller, A. and A.J. Tomarken, Task-dependent changes in frontal brain asymmetry: effects of incentive cues, outcome expectancies and motor responses. Psychophysiology, 2001. 38: p. 500-511.
36. Sobotka, S.S., R.J. Davidson, and J.A. Senulis, Anterior brain electrical asymmetries in response to reward and punishment. Electroencephalography and Clinical Neurophysiology, 1992. 83: p. 236-247.
37. Coan, J.A. and J.J.B. Allen, Frontal EEG asymmetry and the behavioural activation and inhibition systems. Psychophysiology, 2003. 40: p. 106-114.
38. Harmon-Jones, E. and J.J.B. Allen, Behavioural activation sensitivity and resting frontal EEG asymmetry: covariation of putative indicators related to risk for mood disorders. Journal of Abnormal Psychology, 1997. 106(1): p. 159-163.
39. Sutton, S.K. and R.J. Davidson, Prefrontal brain asymmetry: a biological substrate of the behavioural activation and inhibition system. Psychological Science, 1997. 8(3): p. 204-210.
40. Fairclough, S.H. and L. Venables, Prediction of subjective states from psychophysiology: a multivariate approach. Biological Psychology, 2006. 71: p. 100-110.
41. Cacioppo, J.T. and L.G. Tassinary, Inferring psychological significance from physiological signals. American Psychologist, 1990. 45(1): p. 16-28.
42. Coan, J.A. and J.J.B. Allen, Frontal EEG asymmetry as a moderator and mediator of emotion. Biological Psychology, 2004. 67: p. 7-49.
43. Wilson, G.F., J.D. Lambert, and C.A. Russell. Performance Enhancement with Real-Time Physiologically Controlled Adaptive Aiding. in Human Factors and Ergonomics Society Annual Meeting. 2007. Baltimore, Maryland: HFES.
44. Wilson, G.F. and C.A. Russell, Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Human Factors, 2007. 49(6): p. 1005-1018.
45. Pantic, M. and L.J.M. Rothkrantz, Towards an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE, 2003. 91(9): p. 1370-1390.
46. Ahn, H., et al. Stoop to conquer: posture and affect interact to influence computer users' persistence. in Second International Conference on Affective Computing and Intelligent Interaction. 2007. Lisbon, Portugal.
47. Fairclough, S.H., Psychophysiological inference and physiological computing games, in BrainPlay07: Brain-Computer Interfaces and Computer Games. 2007: Salzburg.

Cross-modal Elicitation of Affective Experience

Christian Mühl and Dirk Heylen
Human Media Interaction
University of Twente, NL
c.muehl@utwente.nl

Abstract

In the field of Affective Computing the affective experience (AX) of the user during the interaction with computers is of great interest. Physiological and neurophysiological sensors assess the state of the peripheral and central nervous system. Their analysis can provide information about the state of a user. We introduce an approach to elicit emotions by audiovisual stimuli for the exploration of (neuro-)physiological correlates of affective experience. Thereby we are able to control for the affect-eliciting modality, enabling the study of general and modality-specific correlates of affective responses. We present evidence from self-reports, physiological, and neurophysiological data for the successful induction of the affective experiences aimed for, and thus for the validity of the elicitation approach.

1. Introduction

Affective computing aims at an enrichment of HCI by taking the user's affective state into account [29]. Thereby, applications can unfold their functions in the context of user experience, ideally leading to the increase of the bandwidth and naturalness of interaction.

To achieve such enhanced interactions a robust automatic recognition of the user state is a necessary prerequisite. In the past years the automatic analysis of affect-related behaviours, especially those evident in facial expression or voice, yielded promising results [10, 41]. Still, the classification of affective user state is no trivial endeavor, as the subjective state, the experience of the user, is not necessarily observable by external means as cameras or microphones.

The analysis of physiological responses during affective experience offers an alternative to the analysis of behavioural responses [6–8,20,23,24,30,39]. However, whilst observable behaviour as facial expressions or voice, can be conveniently studied in the field, physiological and neurophysiological responses are less readily available. Therefore the elicitation of affective experience in the laboratory is still a necessary step to acquire physiological and neurophysiological databases. This data can then be analysed in order to extract features capable of discriminating between affective experience. These are then the basis to develop and refine suitable classification methods using those features.

To explore the generalisation of physiological and neurophysiological correlates of affective experiences we developed a cross-modal elicitation method. Specifically, we constructed a set of audiovisual stimuli to be able to elicit emotions either from the auditory or from the visual modality. This study presents evidence, based on the subjects' self-assessments, and on preliminary physiological and neurophysiological results, for the induction of different emotions, and thereby for the validity of the approach.

Before we outline our research questions in more detail, we will introduce the reader to the issue of the validation of emotion elicitation approaches, and to our specific approach.

1.1. The validation of an elicitation method

One can discriminate between endogeneous and exogeneous elicitation methods [30]. The former require the subject to induce affective experiences by remembering or imagening emotional episodes [1, 6, 30, 31]. The latter approach makes use of affective stimuli or tasks to elicitate corresponding experiences. A wide variety of affective stimuli has been used for this purpose, among them pictures [5, 7, 26], naturally occurring sounds [3], music pieces [13, 22, 33], films [15, 19, 23, 39], manipulated applications [20], and computer games [8, 40]. In our approach, outlined below, we will use affective stimuli.

A general problem accompanying the induction of affective experience is the validation of the induction method [14, 38]. Fairclough [14] discusses several methods that can be applied to ensure this *concurrent validity* of the elicitation approach in the context of psychophysiological measurements.

For the use of stimuli or tasks one has to be fairly confident that they indeed induce the target states. The use of

978-1-4244-4799-2/09/$25.00

normed stimulus sets, as the IADS [4] or IAPS [25] can make this more likely. Similarly when using tasks one can use standardized tasks developed within the field of experimental psychology. Alternatively, one might use tasks that have known effects on the user, for example manipulated computer games. However, as Fairclough points out, the use of these latter approaches is close to a natural context, but also prone to confounds due to the complexity of real-world situations.

Another method for the labeling and validation of the data, especially in the domain of facial expression or voice analysis, are observer ratings of the participants behaviour. However, the occurrence frequency of behaviour might be low in the cases that we are considering where the participant is restrained by recording equipment.

Alternatively, one can apply self-assessment methods as the Self-Assessment Manikin [2], to ensure the success of the elicitation method. This, of course, comes for the price of a possible interference with the target behaviour, and an added risk of artifact production. Additionally, self-assessments are not free of bias and dependent on a truthful report.

Furthermore, one might record physiological or neurophysiological data and contrast the different conditions. Should one find a difference between conditions, this can be taken as evidence that indeed different states were induced. To ensure that the desired states were induced one would have to measure and contrast physiological variables that were shown before to vary with the target state or dimension. Those variables, for example, could be chosen by an extensive literature review or the consultation of an expert.

Figure 1. The relationship between stimuli, self-assessment, and physiological data, and the elicitated affective experience.

Figure 1 illustrates the above described relationships between administered stimuli, logged self-assessments, and physiological data. The elicited affective experience can be validated by each of these. However, it has to be mentioned that none of the validation methods is perfect and thus a combination of different validation procedures might yield the most insight into the concurrent validity of the elicitation approach.

In this paper we will analyse the results of the first series of elicitation experiments that we carried out with our approach. We look at how self-assessments relate to the affective states that we intend to elicit and to what extent (neuro-)physiological measures can discriminate between the various stimuli groupings.

1.2. The cross-modal elicitation of affective experiences

As already outlined above, physiological and neurophysiological signals carry information about the affective state of the user. While relatively many studies explored the potential of physiological features to differentiate affective experiences [8,20,23,24,30,39], only few studies looked at the suitability of neurophysiological sensors [6, 7]. Most studies were conducted under controlled circumstances. This is especially true for the EEG studies. The tight control of experimental protocols is a necessary prerequisite to disentangle the manifold physiological processes occurring in real-world environments, and thus to avoid confounding variables. However, it is also impeding the ecological validity of the psychophysiological inferences made on the basis of such simple elicitation paradigms [14]. To ensure the generalisation of the feature-experience relationships found in the controlled laboratory experiments to real-world applications, a slow increase in the complexity of the experiments and finally the step into the field seems advisable.

Our motivation for the development of the elicitation method introduced here is to make a modest step in this direction, exploring the generalisation of psychophysiological inferences over different affect elicitation modalities, but still staying inside the laboratory. We are aiming for the controlled induction of affective experiences via the visual or auditory stimulus modality. Specifically, we want to manipulate experience along the valence dimension of the dimensional emotion model according to Russel [32], that is to elicitate negative, neutral, and positive affective experiences. For this purpose we combine affective neutral and valence-carrying stimuli from different unimodal stimulus sets to a new multimodal stimulus set.

We chose sound (IADS) and picture (IAPS) stimulus sets to construct new audiovisual stimuli. One advantage of those specific sets is that they are normed by hundreds of participants according to their effect on the participants' affective experience. The knowledge about mean valence and arousal responses for a given stimulus guides our combinations of neutral and valence-carrying unimodal stimuli to one multimodal. However, one should take in mind the standard deviations of the norm ratings are quite big and

show a large spread of responses from different subjects to a given stimulus over the arousal-valence plane. This indicates a subject- and context-specific response to the stimuli. Furthermore, to construct our stimulus set we make combinations of different stimuli from the original databases. It cannot be assumed that the combination of different affective stimuli has a linear effect on the affective response. The combination of picture and sound might produce a different context, changing or even inverting the original affective response caused by the valence-carrying stimulus. Despite our intentions to control for that by a careful choice of combinations of auditory and visual stimulus parts, the elicitation of the target emotions has to be shown to ensure the validity of our elicitation method. In the following section we will describe the construction of the new multimodal stimulus set in detail.

1.3. Stimuli construction

To study the effects that the different modalities have on neurophysiological affective responses, 180 multimodal stimuli were constructed from the auditory and visual affective stimuli sets IADS and IAPS.

From each stimulus set, IADS and IAPS, we chose 30 stimuli from the positive and 30 stimuli from the negative side of the valence dimension. Additionally, we chose 60 neutral auditory and 60 neutral visual stimuli from each modality. Note that we employed each neutral unimodal stimulus twice (due to the low number of IAPS stimuli). Each neutral stimulus from one data set would appear one time in combination with another neutral stimulus from the other data set and one time in combination with a valence-carrying stimulus of the other data set. The three different valence intervals, positive, neutral, and negative, were defined according to the mean ratings on the valence scales. The 9-point valence Likert scale the norm-ratings are based on are ranging from 1(feeling unhappy) to 9 (feeling happy). Therefore, we required positive stimuli to have a mean rating above 6.5, negative stimuli to have a mean rating below 3.5, and neutral stimuli to lie in between these two groups.

We constructed five groups of auditory-visual stimuli: (1) *auditory negative*, (2) *auditory positive*, (3) *visual negative*, (4) *visual positive*, and (5) stimuli that were neutral both auditory and visually, referred to as *multimodal neutral*. An auditory negative stimulus consisted of a negative auditory stimulus and a neutral visual stimulus. An auditory positive stimulus contained a positive auditory and neutral visual stimulus. This way the affect elicitation was supposed to result from the auditory stimulus. Correspondingly, the visual negative and positive stimuli were created from a neutral auditory and a valence-holding visual stimulus. The multimodal neutral stimuli consisted of a neutral auditory and a neutral visual stimulus. This group was important as a control condition, which enables the analysis of the specific effects of positive and negative stimulation, respectively. While the grouping was based on the distribution of the stimuli on the valence axis, we tried to keep the group differences on the arousal axis comparable to avoid confounding effects. Specifically, we tried only to use stimuli that had a relatively high arousal value, i.e. higher than 3.5. Because of a bias in the original sets, we were not able to do this.

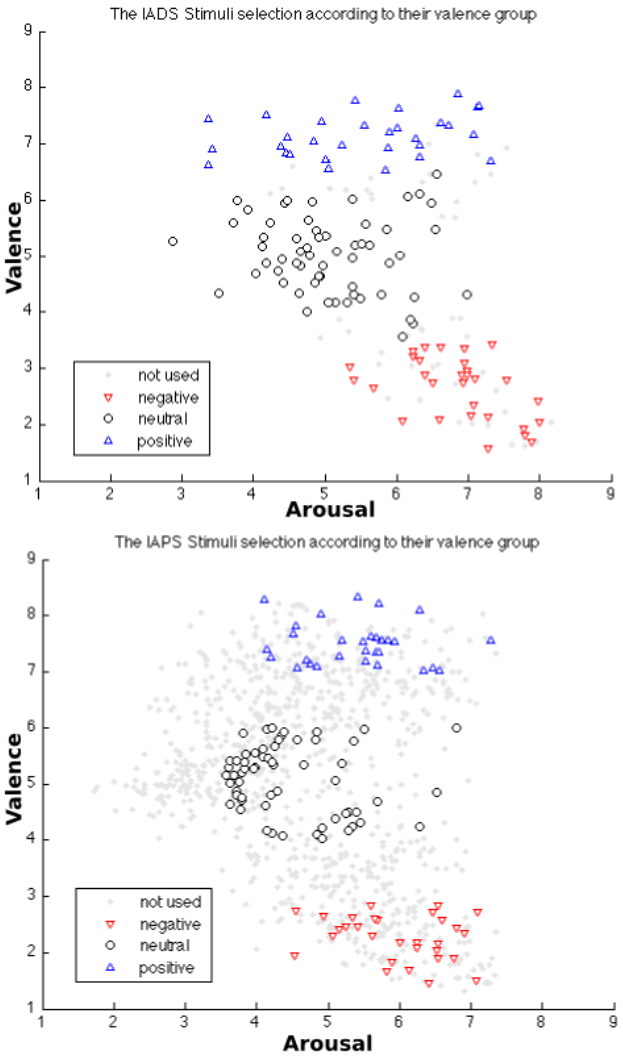


Figure 2. The position of the selected auditory (above) and visual (below) stimuli in the valence-arousal space.

As described above, the stimuli were constructed from carefully selected subsets of the IADS and IAPS. Figure 2 shows the positive, neutral, and negative stimuli groups chosen from the IADS and IAPS sets, and their locations in the valence-arousal plane. The stimulus sets used for our stimulus construction are not evenly distributed in the valence-arousal plane, but describe a C-form. There are almost no

stimuli that have low arousal and high valence values, or high arousal and medium valence. On the other hand are the negative stimuli in general more arousing than positive or neutral stimuli. In consequence a selection of stimuli that differ in valence, but not in arousal is only possible to a limited degree. The higher the number of stimuli is, the stronger is the selection influenced by the mentioned IAPS and IADS characteristics. The most limiting factor is the small size of the IADS, which makes it difficult to select more than 30 negative and 30 positive stimuli.

Group	Valence mean (std)	Arousal mean (std)
A positive	7.14 (0.38)	5.51 (1.16)
A neutral	5.01 (0.67)	5.06 (0.85)
A negative	2.65 (0.55)	6.84 (0.71)
V positive	7.49 (0.38)	5.40 (0.77)
V neutral	5.08 (0.60)	4.46 (0.77)
V negative	2.27 (0.40)	5.97 (0.73)

Table 1. The mean valence and arousal ratings per modality and stimulus group.The value in brackets is the standard deviation.

Table 1 gives an overview over the group's valence and arousal means according to the norm ratings from the IADS and IAPS manuals. The valence means of the groups are all significant different. Despite our efforts to keep the arousal equal over the groups, also the arousal means are significantly different. However, as the norm values of the IAPS and IADS are already characterised by a big standard deviation, it was not assumed to be able to predict the precise effect of the stimuli on a particular group of participants. In that respect the valence and arousal values were only used as an initial strategy to select the optimal stimuli for our purpose.

1.4. Research questions

To validate our elicitation approach, we are interested in the effect that our stimulation has on the participant's experience. As described above there are different strategies that one can use to ensure that the affective experience of interest was induced. Therefore, we analysed the participants' self-assessments according to the different stimuli categories employed, irrespective of the elicitation modality. Furthermore, we analysed the (neuro-)physiological responses to the different stimulus categories employed. Finally, we explored the effect of the choice of an alternative ground truth, that is the (neuro-)physiological responses to different groupings of the trials according to the self-assessments of the subjects.

Our main question was whether the target emotion is indeed induced by our elicitation paradigm. We expected that in the comparison of the self-assessments given after each stimulus presentation, the valence judgements over stimuli and subjects would be significantly different between conditions. On the other hand, arousal should ideally be comparable, as we aimed for similar arousal values during the construction of the stimulus groups.

A further expectation was that the comparison of the physiological responses during the presentations of the different stimulus groups yields significant differences. Especially, we expected differences for those physiological and neurophysiological sensors implied before in valence-related nervous system responses. Physiological correlates of valence manipulations have been found for electromyographical responses recorded from the facial muscles (EMG) [5, 39], in electrocardiographical recordings (ECG), specifically heart rate [31, 33], and for blood pressure [36]. Neurophysiological correlates, specifically those derived from electroencephalography (EEG), include the asymmetry of alpha power between the left and right hemisphere of the brain [11], and frontomedial theta power [33].

Significant differences in other (neuro-)physiological signals not directly implied in valence-, but other affect-related experiments might also offer evidence about different states induced, though they could not be used as evidence for the elicitation of the target states. Physiological signals implied in arousal-related manipulation of affective experiences are the galvanic skin response (GSR) [3,9,22,26], the respiratory sinus arrhythmia (RSA) derived from the heart rate [15], and respiration [13, 17]. Neurophysiological arousal-specific responses include a decrease of the overall level of power in the alpha band [27] and an increase of power in the gamma band [21, 28].

Finally, the question most relevant for the determination of a ground truth for later classification approaches is, if there is a more favourable grouping (of trials into conditions) possible according to the self-assessments. That there is a significant difference between classification results achieved via a norm based and a self-assessment based ground truth was shown by Chanel et al. [7]. Therefore, we resorted the stimuli, and thus the trials, into positive, neutral, and negative affect conditions according to self-assessment. The trivial assumption was that this regrouping would create more homogeneous conditions, with condition means further apart, and smaller standard deviations. Furthermore, we expected more significant (neuro-)physiological differences, especially for sensors implied in valence manipulation before.

2. Methods

2.1. Participants

14 participants (7 men and 7 women) took part in the experiment. Due to incomplete recordings the data of two participants was not analysed. The participants were aged between 19 and 53 (mean age 28) and all except one indi-

cated their right hand as the dominant hand.

2.2. Stimuli

For the experiments the newly constructed audiovisual stimulus set as described in section 1.3 was used. To avoid eye-movements during the stimulus presentations the pictures were decreased in size to 400 x 300 pixels. Primary stimulus characteristics as overall loudness of sounds or brightness of visual stimuli may have significant effects on neurophysiological data. To minimize the risk of a confound by stimulus-related non-affective characteristics we tested the group differences of mean subjective loudness and mean luminance. No significant differences between the visual parts of the positive, negative, and neutral group was found in terms of mean luminance. Similarly, no significant differences between the auditory parts of the positive, negative, and neutral group in terms of mean subjective loudness could be detected.

2.3. Equipment and signal acquisition

2.3.1 Presentation and recording hardware

The stimuli were presented on a dedicated stimulus PC (P4 3.2GHz), which sent markers according to stimulus on- and offset to the EEG system (Biosemi ActiveTwo system, www.biosemi.com). For the stimulus presentation we used "Presentation" (Neurobehavioral systems, www.neurobs.com). The visual parts of the stimuli were presented on a 20 inch monitor (Samsung SyncMaster 203B). The auditory parts of the stimuli were presented via a pair of custom computer speakers (Phillips Multimedia Speaker System). The distance between participants and monitor/speakers was about 70 cm.

The physiological and neurophysiological signals were recorded with 512 Hz on a dedicated recording PC (P4 3.2GHz) running Actiview software (BioSemi).

We recorded from 64 active silver-chloride electrodes placed according to the the 10-20 system. Additionally, 4 electrodes were applied to the outer canti of the eyes and above and below the right eye to derive horizontal EOG and vertical EOG, respectively.

Besides recording neurophysiological signals by electroencephalography we assessed also the state of the peripheral nervous system via several physiological sensors.

To obtain the electrocardiogram we placed an electrode at the inner side of the left arm of the participant. A plethysmograph was clipped to the left index finger to assess blood volume pulse. A temperature sensor was placed on the distal phalange of the small finger of the left hand to measure peripheral temperature. Respiration was assessed via a respiration belt placed around the chest just over the stomach. To assess the activity of the somatic nervous system we applied electrodes to two facial muscles, the right corrugator supercilii (implied in frowning) and the left zychomatic major (implied in smiling). The EMG sensor placement over the zygmaticus major and the corrugator supercilii muscle was done via two electrodes for each muscle and according to the guidelines from [16] on the left cheek and over the right brow, respectively.

2.4. Procedure

The Participants were seated in a comfortable chair in front of monitor and speakers. They read an informed consent form and user instructions before the experiment. After filling in a questionnaire and signing the informed consent the EEG cap and the physiological sensors were placed according to the descriptions above. Before the start of the experiment the participant was introduced to the Actiview online view of her EEG signals to make her conscious of the influence of movement artifacts. She was instructed to restrict movements to the periods between trials. Finally, the SAM scales were explained, so that a good understanding of the concepts of arousal and valence could be assured. Participants were advised to give a "gut response" to emphasise the importance of their subjective feeling and to avoid a more cognitive judgement of the stimuli themselves.

2.5. Experiment Design

The stimulus presentation was done in 4 blocks with 45 stimuli each. The order of the stimuli presentation was randomised for each participant. To avoid tensions or fatigue, in the breaks the participant could correct seating position, drink, and relax until she felt ready to continue. Figure 3 depicts the trial structure employed. Below we will outline each of the trial periods and its functions.

Pre-stimulus phase Two seconds before a stimulus is presented a fixation cross is blended into the middle of the screen. This cross is supposed to limit eye movement during stimulus presentation and will be kept on the screen until the self-assessment phase.

Stimulus phase The stimulus is presented for six seconds, which is the length of the auditory stimulus. The visual counterpart is shown during the time the sound is played.

Post-stimulus phase Between the stimulus offset and the begin of the self-assessment the fixation cross is further visible on a black background for two seconds. This phase is intended to serve as a stimulus free period in which the independence of a potential affect-related neurophysiological response from the stimulus characteristics can be shown.

Figure 3. Example trial with the six trial periods and their duration (arrows indicate self-paced rating phases).

Self-assessment phase The norm ratings of the IAPS and IADS are characterised by a considerable variance per stimulus. Thus a given stimulus might induce different affective states in different subjects. To study the effectiveness of our affective stimulation and to explore alternative groupings for the signal samples in positive, neutral and negative trials, a self-report in form of the self assessment manikin (SAM; see [2]) is employed after each stimulus presentation. The duration of the rating phases for arousal and evaluation is not limited. It ends as soon as the user finishes the self-assessment. However, the subject is instructed to answer by a fast intuitive judgement.

Resting phase The physiological response is known to be relatively slow, peaking around five seconds after stimulus presentation [37]. To reduce the contamination of the samples by prior samples, the rating is followed by an interstimulus interval of averagely five seconds. The participants are also instructed to blink and move preferably in this period, to decrease the contamination of the trials by movement artifacts.

2.6. Preprocessing of EEG data

We used EEGlab [12] to preprocess the EEG data. Specificly, we computed the common average reference (CAR), downsampled the data to 256 Hz, and high passed it with an infinite impulse response filter at 1 Hz. Then we extracted epochs of six seconds, from stimulus onset to stimulus offset. We computed the absolute frequencies for the theta (θ, 4 - 7 Hz), alpha (α, 9 - 12 Hz) via a FFT with a sliding window length of 128 samples and 50% overlap.

Furthermore, we computed the asymmetry for each pair of the left and right frontal channels, that is AF3 and AF4, and F3 and F4, and F5 and F6, in the alpha frequency band by formula 1.

$$X_{asym} = \frac{(X_{left} - X_{right})}{(X_{left} + X_{right})} \quad (1)$$

As we did not remove potential artifacts from the data, we restricted our analysis to the alpha and theta frequency bands. Furthermore, we focused on the analysis of anterior regions of interest, as we expected modality-related variations in EEG power in the posterior modality-specific regions. Figure 4 shows the electrode layout for the frontal cortex. We extracted the power of the alpha band for the left and right frontal regions, and the power of the theta band for the fronto-medial region.

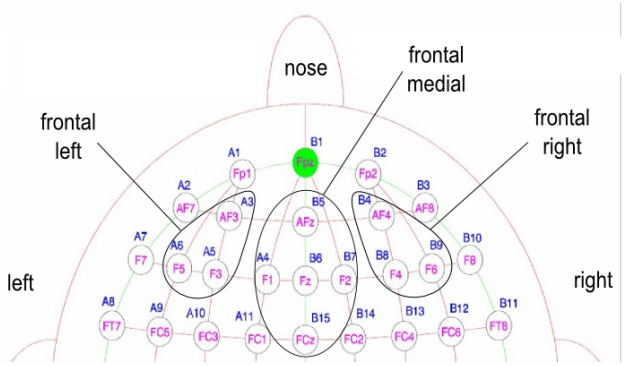


Figure 4. The regions of interest for the preliminary analysis of EEG signals.

2.7. Preprocessing of physiological data

As most physiological sensors are known to have a slow response to stimulation and thus a long response latency, we extracted long epochs of ten seconds for each trial. An epoch contained the pre-stimulus period, the stimulation period and the post-stimulus period. From the signal part that contained the stimulus and post-stimulus period we obtained several features for each of the measured biosignals, while the pre-stimulus part of the signal was used for baseline removal for sensors that are susceptible to stimulus-independent long-term variations. We sampled the physiological signals down to 256 Hz. Below we describe the extracted features for the cardiovascular signals, the galvanic skin response, and the facial EMG sensors in detail.

Cardiovascular features In Table 2 the extracted cardiovascular features are described. For the extraction of the heart beats and the computation of the highest frequency of the heart rate variability, the respiratory sinus arythmia, the BIOSIG toolbox for Matlab was used ([35]). To eliminate the effect of stimulus-independent, low frequency fluctuations in the blood volume pulse data, we subtracted the baseline mean from each trial.

Feature	Description
$E\{h\}$	mean heart rate
HF	highest frequency of the heart rate variability
$E\{b\}$	mean of the blood volume pulse
$\sigma\{b\}$	standard deviation of the blood volume pulse
$min\{b\}$	minimum of the blood volume pulse
$max\{b\}$	maximum of the blood volume pulse
$\delta^b_{\|1\|}$	mean of the abs. of the 1. difference of BVP
$\delta^b_{\|2\|}$	mean of the abs. of the 2. difference of BVP
$E\{t\}$	mean T
$\sigma\{t\}$	standard deviation of T

Table 2. The cardiovascular features derived from the electrocardiogram (ECG), blood volume pulse (BVP) and skin temperature (T) sensors and their description.

Galvanic skin response To analyse the galvanic skin response we first low-pass filtered the signal at 0.05 Hz via an infinite impulse response filter of length 4. To further reduce the stimulus independent variance of the data, we de-trended each trial and subtracted the baseline mean. Table 3 shows the features extracted from the filtered signal.

Feature	Description
$E\{s\}$	mean skin conductance
$\sigma\{s\}$	standard deviation of the SC
$\delta^s_{\|1\|}$	mean of the abs. of the 1. difference of SC
$\delta^s_{\|2\|}$	mean of the abs. of the 2. difference of SC

Table 3. The features derived from skin conductance sensors (SC).

Facial electromyography According to [39] the two electrode pairs placed over the right corrugator supercilii and the left zychomatic major were subtracted, yielding the EMG signals for each muscle, from which we extracted the first four statistical moments, as enlisted in Table 4.

3. Results

In a preliminary analysis we studied the recorded data to gain insights into the validity of our approach. Furthermore we hoped to learn which grouping method, according to stimulus norms or according to self-assessment, would be better suited as ground truth for future in-depth study of the physiological and neurophysiological correlates. We first will present the self-assessment data for the different grouping methods. Then, we will examine the physiological and neurophysiological differences between the 3 conditions, positive, negative, and neutral emotions, for the different grouping methods.

Feature	Description
$E\{c\}$	mean CS
$\sigma\{c\}$	standard deviation of the CS
$kurt\{c\}$	kurtosis of the CS
$skew\{c\}$	skewness of the CS
$E\{z\}$	mean ZM
$\sigma\{z\}$	standard deviation of the ZM
$kurt\{z\}$	kurtosis of the ZM
$skew\{z\}$	skewness of the ZM

Table 4. The EMG features derived from the right corrugator supercilii (CS) and the left zychomatic major (ZM).

3.1. Analysis of the self-assessment data

The evaluation of the self-assessment is not only a mean to validate our emotion induction method, but also gives us the possibility for an alternative grouping of the stimuli according to the individual affective response toward each multimodal stimulus. The grouping of the stimuli establishes the ground truth in the search for physiological and neurophysiological differences and for a later classifier training.

The analysis of the mean stimulus valences suggested that different stimulus groups (positive, neutral, negative) resulted in different affective experiences. The mean values behaved according to the group membership. However, for many stimuli the induced emotions differed from the emotions the stimuli were supposed to induce. This was also reflected by participants informal reports after the experiments. For example, a starving child on a blue blanket was perceived by one participant as cared for and elicited a calm and rather positive response, while it was intended to elicit a negative reaction. Figure 5 shows the distribution of valence and arousal ratings over all stimuli and subjects for the five conditions, also taking the modality of the affect eliciting stimulus into account. Despite the clear differences that are visible between the groups, the distributions are overlapping to a large degree. That is, some of the stimuli had not the intended effect on some subjects, but instead elicited another affective state.

These deviations of the individual affective experience from the target affective experience of the stimuli are natural taking the individual differences between participants

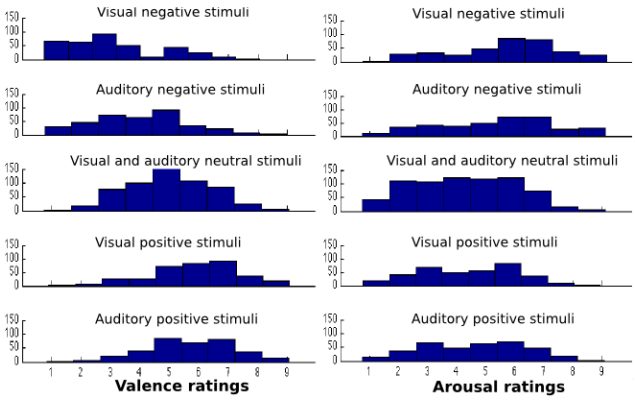


Figure 5. The distributions of the SAM ratings for the original groupings for valence (left) and arousal (right).

	N+	Nn	N-	sum
S1+	270	104	38	412
S1n	386	519	313	1218
S1-	60	97	369	526
sum	716	720	720	2156
S2+	426	216	92	734
S2n	162	308	140	610
S2-	128	196	488	812
sum	716	720	720	2156

Table 5. The contingency table shows the relationship of the stimulus grouping into the affect conditions (+ = positive, n = neutral, - = negative stimuli) according to the IADS and IAPS stimuli norms (N group) and to the self-assessment with normal-sized (S1 group in upper table) and small-sized (S2 group in lower table) neutral condition.

into account. They are already reflected in the variances that characterise the norm ratings of the individual stimuli in the IAPS and IADS. Therefore, the question arises, if another sorting of the trials into the positive, neutral, and negative experience conditions could result in more homogeneous conditions of affective experiences. This is especially interesting for a later use of the data for the identification of physiological and neurophysiological features able to differentiate between the affective states.

To explore the potential of alternative ways to assign the trials to the conditions we made use of the gathered self-assessment data. We will refer to the alternative ways of sorted trials into conditions as grouping approaches, resulting in different groupings of the trials.

The most obvious grouping approach, and the one used so far, is to sort the recorded trials according the norm ratings of the affect-eliciting stimuli, coinciding with the stimuli conditions we constructed. In the following we will refer to this grouping approach as NORM, or as N in the tables.

When the self-assessment values are used as the basis for the stimuli- and thus trial-grouping, we obtain the SAM1 grouping (S1 in the tables). The deviations from the intended grouping become obvious, when directly comparing the overall number of trials per condition intended (positive: N+, neutral: Nn, negative: N-) with those derived from the new grouping approach (S1+, S1n, S1-) in a contingency table 5. Due to a rating trend towards the middle, thus toward the neutral condition, the positive and negative conditions are underrepresented in the number of trials.

However, by assuming each rating that deviates from the middle of the Likert scale by one scale unit towards one end of the scale to result from a negative or positive affective response, we obtain the SAM2 grouping (S2 in the tables). Here the responses are equally distributed over all three conditions (S2+, S2n, S2-), as the neutral condition is narrowed down to one Likert point. The relationship between the intended grouping, by use of IAPS and IADS norm values, and this this less strict grouping due to self-assessment can again be seen in table 5.

Group	Valence mean (std)	Arousal mean (std)
N+	5.29 (1.58)	4.01 (1.84)
Nn	4.49 (1.35)	3.75 (1.86)
N-	3.04 (1.70)	5.02 (2.03)
S1+	6.79 (0.64)	3.71 (1.94)
S1n	4.4 (0.74)	3.76 (1.78)
S1-	1.79 (0.75)	5.88 (1.59)
S2+	6.22 (0.81)	3.79 (1.89)
S2n	4.47 (0.13)	3.41 (1.70)
S2-	2.37 (0.98)	5.34 (1.79)

Table 6. The mean valence and arousal ratings per group and grouping method. The value in brackets is the standard deviation.

Table 6 enlists the means and standard deviations for the positive, neutral, and negative conditions according to the different grouping methods. The grouping of the stimuli based on the self-assessment leads to a clearer distinction of the conditions in terms of valence means and to a smaller standard deviation. As a consequence of the SAM2 grouping variation, however, the differences between condition means are decreasing again and the standard deviations of positive and negative condition are increasing. A Wilcoxon signed-rank test on the valence ratings revealed statistical significant differences ($p \leq 0.001$) for all emotion contrasts within all three grouping approaches. The same was observed for the arousal ratings, except for the contrasts of positive and neutral conditions. These differences were due to a higher arousal induced by the negative stimuli.

Summarising, the analysis of self-assessment rating means of the conditions of the NORM suggests that indeed different affective experiences were elicited. Further-

more, it was shown that the alternative grouping according to the self-assessments, to make the conditions more homogeneous in terms of elicited emotion, results in an imbalance of trials per conditions. This can be remedied by the limitation of the neutral condition to those trials that were not accompanied in a deviation from the central, and thus most neutral bin, of the self-assessment valence scale. Furthermore, we found differences in the arousal dimension, which have to be taken into account in the further study of the data.

To explore the effect of the different grouping methods in terms of physiological and neurophysiological differences, we analysed a subset of the available sensor information.

3.2. Analysis of the physiological data

We conducted a preliminary analysis of the physiological signals according to the NORM, SAM1 and SAM2 grouping. We used the non-parametric Wilcoxon signed-rank test to test for differences between the extracted features, as some of the groups were not normally distributed. The features shown in table 7 are significant with a p-value ≤ 0.05. (For this preliminary analysis we did not correct for the multiple tests conducted.)

Contrast	Significant Features
N+ vs N-	HF , $\sigma\{b\}$, $E\{t\}$
N+ vs Nn	$E\{h\}$, $E\{s\}$,$\sigma\{s\}$
N- vs Nn	$E\{h\}$, $E\{s\}$, $\sigma\{s\}$, $E\{t\}$
S1+ vs S1-	
S1+ vs S1n	$\sigma\{s\}$, $\delta^s_{\lvert 1\rvert}$, $\sigma\{z\}$
S1- vs S1n	$\sigma\{c\}$
S2+ vs S2-	$\sigma\{z\}$
S2+ vs S2n	HF , $\delta^s_{\lvert 1\rvert}$, $\sigma\{z\}$
S2- vs S2n	

Table 7. The significant ($p \leq 0.05$) physiological features for the contrasts of negative (-), neutral (n), and positive (+) stimulus groups according to the NORM (N), SAM1 (S1), and SAM2 (S2) grouping methods.

Surprisingly, the NORM grouping results in the most significant differences between the conditions. Heart rate, Heart rate variability, blood volume pulse, temperature, and skin conductance are differentiating between the conditions. Specifically heart rate and skin conductance seem to differ between the emotional and the neutral conditions, while heart rate variability, blood volume pulse and temperature are differentiating the two valenced conditions.

For the SAM1 and SAM2 grouping we observed only differences in skin conductance, heart rate variability, and muscle activity. Intriguingly, the corrugator supercilii muscle (implied in frowning) is differentiating the negative condition from the neutral condition in the SAM1 grouping, while the zychomatic major muscle (implied in smiling) is differentiating between positive and neutral condition for both SAM groupings. Unexpectedly, two of the SAM contrasts could not be differentiated in terms of physiological responses.

3.3. Analysis of the neurophysiological data

For the preliminary analysis of the neurophysiological sensors according to the NORM, SAM1 and SAM2 grouping we concentrated on the alpha and theta frequency over the lateral and medial frontal cortex. Again we used the non-parametric Wilcoxon signed-rank test, as some of the groups were not normally distributed. Table 8 shows the significant ($p \leq 0.05$) features. (As in the previous analysis of physiological features we did not correct for the multiple tests conducted.)

Contrast	Left α	Right α	Medial Θ
N+ vs N-		AF4	
N+ vs Nn			
N- vs Nn			
S1+ vs S1-	AF3		
S1+ vs S1n	AF3, F5	AF4, F4, F6	FCz
S1- vs S1n			FCz
S2+ vs S2-	F3, F5		
S2+ vs S2n	F3, F5		
S2- vs S2n			FCz

Table 8. The significant ($p \leq 0.05$) EEG features for the contrasts of negative (-), neutral (n), and positive (+) stimulus groups according to the NORM (N), SAM1 (S1), and SAM2 (S2) grouping methods.

The most salient finding is the lower alpha power for the positive conditions. As there is a reciprocal relationship between alpha power and neural activity, this might indicate a stronger processing of positive stimuli.

The tests for alpha asymmetry between the electrode pairs AF3 and AF4, and F3 and F4, and F5 and F6 showed no significant differences.

Higher fronto-medial power in the theta band was found for neutral compared to negative conditions in the self-assessment contrasts. This relates to the study of Sammler et al. [33]. They found a fronto-medial increase in theta power for normal (positive) compared to distorted (negative) musical pieces and interpreted it as an emotional reaction associated with attentional processes. However, for the SAM1 grouping we found a decrease of theta power for positive compared to neutral trials, which seems to be a contradiction. A reconciliation is possible if one assumes that emotional stimuli in general might trigger these attentional processes observed over fronto-medial cortices.

Similar to the analysis of the physiological features we

see the biggest difference between the normed grouping method on the one side and the two self-assessment based groupings on the other side. However, in contrast to the previous analysis, we now see the strongest difference for the self-assessment groupings, especially for SAM1.

4. Discussion

The analysis of the self-assessment data provided evidence for the validity of our stimulus sets. However, we observed a great variability in valence ratings for a given stimulus over subjects. This was to a certain degree expected, as individual differences already caused large variations in the ratings of the original stimulus sets of IADS and IAPS. We used multimodal stimuli, which were constructed of a valenced and a neutral unimodal stimulus. This might have weakened the effectiveness of the used stimuli further, leading to the observed trend of ratings towards the middle, i.e. the neutral condition.

Furthermore, we found significantly higher mean arousal values for the negative stimuli in the self-assessment data. This reflected the arousal bias observed in the norm ratings of the negative stimuli subsets. This effect has to be taken into account, when the (neuro-)physiological correlates of the affective experience elicited by the negative stimuli are interpreted, as a difference solely due to the experience difference in the valence dimension cannot be ensured.

As the choice of the right ground truth, the sorting of trials to the affective experience conditions, can have significant consequences for later classification attempts, we explored three ways to sort the recorded trials into positive, neutral, and negative conditions. The grouping of the trials according to the constructed stimuli groups (NORM grouping), exhibited large standard deviations, resulting from those stimuli that did not have the expected effect on the participants. To build more homogeneous conditions in terms of self-assessment ratings we grouped the trials according to those ratings (SAM1 grouping). Due to individual differences in rating styles this led to imbalanced group sizes. Specifically the negative and positive conditions contained only a small number of trials relative to the neutral condition. By a limitation of the neutral condition to the most central bin on the rating scale, we achieved a more balanced distribution (SAM2 grouping). However, also the self-assessment method is not free from biases or distortions [34]. Therefore, it is not necessarily the optimal choice for a solid ground truth construction.

A fourth sorting alternative would be a combination of stimulus reliability across participants and individual self-assessment. That is, to choose only those trials for an self-assessment based analysis, in which stimuli with relative unequivocal ratings were presented. That way we would reduce the overall number of trials, but could avoid the analysis of responses towards stimuli which might induce mixed emotions. These mixed emotions might have led to the variations of ratings for a given stimulus over subjects. By the removal of those stimuli from the data sets more homogeneous conditions could be created.

Another possibility to find suited sets of positive, neutral, and negative stimuli to build a valid ground truth is the use of physiological responses for verification. Marosi et al. [27] analysed only those trials in terms of EEG frequency activity, which were accompanied by a galvanic skin response. However, the analysis of EEG data requires a great amount of trials due to an inherent low signal-to-noise ratio. To explore the feasibility of such an approach the number of those physiological responses in the physiological data has to be determined. Furthermore, such an analysis might only find differences between trials that would theoretically be differentiable by physiological sensors, neglecting EEG features that might also differ between affective experience in the absence of physiological responses.

The preliminary analysis of physiological and neurophysiological sensors gave further evidence for a successful elicitation of different affective experiences by our approach. However, these findings were not free of contradictions. We found large differences between the NORM and the SAM grouping methods in number and type of physiological signals differing between affect conditions. We expected to find stronger differences between the conditions when grouping according to the self-assessments, as reported by Chanel and colleagues [7]. Similar to the current study, they elicited affective states (low, medium, and high arousal) via the presentation of IAPS stimuli. As we did not attempt a classification in the current analysis phase, we cannot directly compare our observed differences with their classification accuracy. However, while Chanel and his colleagues findings indicate a less robust pattern for the norm based grouping in general, we find more physiological features differentiating between conditions in the norm based grouping than in the self-assessment based grouping. For two of the SAM contrasts (S1+ vs. S1- and S2- vs. S2n) we couldn't show any significant effects for the physiological sensors at all. On the other hand, our finding that neurophysiological features do mostly differ between the self-assessment based conditions corroborates the results of Chanel et al. Here the most differences were found for the SAM1 grouping. Although we did not find the expected pattern in terms of alpha asymmetry, we observed consistent decreases of left-hemispheric alpha for the positive compared to the neutral and negative conditions and of fronto-medial theta power for the negative compared to the neutral condition.

The higher number of differences found in the EEG data for the SAM1 grouping could indicate that the emotional responses are more homogeneous for the groups established in this way. However, it might also be the result of some rel-

atively small positive or negative groups, i.e. a small number of samples for some subjects in which possible outliers have a big effect in the statistical analysis. On the other hand, the SAM2 grouping might lead to the inclusion of neutral trials into the positive and negative conditions, and thus obscure the differences between the conditions.

Furthermore, it seems that for the analysis of temporally limited processes an analysis in shorter time windows is important. Didier et al. [18] showed that different subprocesses associated with affective responses are unfolding over different intervals of only few hundred milliseconds in the EEG. However, as auditory stimuli might have big interstimuli variations in the onset of affective response such a division of the trial into subtrials could lead to the comparison of different, unrelated parts of the emotional responses. These inadequate comparisons could lead to further variance in the signals and thus conceal the neural correlates of the emotional processes.

A further exploration of the data is needed to confirm the here presented preliminary results, resolve the contradictions, and find a reliable grouping method for the ground truth construction. The removal of artifacts will give a better insight into the true sources of the physiological and neurophysiological differences between the conditions.

5. Conclusion

We presented an analysis of an emotion elicitation experiment using multimodal stimuli and showed the validity of the experimental approach used along several dimensions. The approach will be used to study physiological and neurophysiological responses associated with affective experience while controlling the emotion-eliciting modality.

The analysis of the self-assessments of the participants emotional states in terms of valence and arousal suggested that the approach used is suitable for the induction of different affective states. However, it was also shown that the variance of the individual responses to the affective stimuli poses a great challenge in the search for (neuro-)physiological correlates of affective processes and their subsequent classification.

We studied different grouping methods to sort the acquired physiological and neurophysiological signals, that is according to the original grouping of stimuli, to the self-assessment data from the valence dimension, and to a self-assessment data with a more relaxed criterion for positive and negative conditions.

The comparison of the physiological features between the three emotion conditions for all three grouping methods revealed a variation of differentiating features over these approaches. Especially the grouping according to the normed values of the used stimuli differed from the two self-assessment groupings in number and types of distinguishing features. The analysis of EEG alpha and theta power revealed a contradictory pattern, with the self-assessment based grouping leading to the best differentiation between conditions.

A further analysis of the neurophysiological and physiological features, incorporating artifact removal and the rejection of particularly unreliable stimuli will yield a better understanding of apparently contradicting phenomena observed in this study. Finally, it will be the first step to an informed choice of features for the exploration of a multimodal affect classification.

Acknowledgements The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

[1] M. Benovoy, J. Deitcher, and J. Cooperstock. Biosignals analysis and its application in a performance setting: Towards the development of an emotional-imaging generator. In *IEEE International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS)*, 2007.

[2] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[3] M. M. Bradley and P. J. Lang. Affective reactions to acoustic stimuli. *Psychophysiology*, 37(2):204–215, 2000.

[4] M. M. Bradley and P. J. Lang. The international affective digitized sounds (2nd edition; IADS-2): Affective ratings of sounds and instruction manual. Technical report, Gainesville: University of Florida, Center for Research in Psychophysiology, 2007.

[5] J. T. Cacioppo, R. E. Petty, M. E. Losch, and H. S. Kim. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50(2):260–268, 1986.

[6] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[7] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Multimedia Content Representation, Classification and Security*, pages 530–537, 2006.

[8] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *MindTrek '08: Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, pages 13–17, New York, NY, USA, 2008. ACM.

[9] M. Codispoti, M. M. Bradley, and P. J. Lang. Affective reactions to briefly presented pictures. *Psychophysiology*, pages 474–478, 2001.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[11] R. J. Davidson. Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1):125–151, 1992.

[12] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[13] J. Etzel, E. Johnsen, J. Dickerson, D. Tranel, and R. Adolphs. Cardiovascular and respiratory responses during musical mood induction. *International Journal of Psychophysiology*, 61(1):57–69, 2006.

[14] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[15] T. W. Frazier, M. E. Strauss, and S. R. Steinhauer. Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology*, 41(1):75–83, 2004.

[16] A. J. Fridlund and J. T. Cacioppo. Guidelines for human electromyographic research. *Psychophysiology*, 23(5):567–589, 1986.

[17] P. Gomez, S. Shafy, and B. Danuser. Respiration, metabolic balance, and attention in affective picture processing? *Biological Psychology*, 78(2):138–149, 2008.

[18] D. Grandjean and K. R. Scherer. Unpacking the cognitive architecture of emotion processes. *Emotion*, 8(3):341–351, 2008.

[19] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[20] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[21] A. Keil, M. M. Müller, T. Gruber, C. Wienbruch, M. Stolarova, and T. Elbert. Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical Neurophysiology*, 112(11):2057–2068, 2001.

[22] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon. Event-related skin conductance responses to musical emotions in humans. *Neuroscience letters*, 328(2):145–149, 2002.

[23] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[24] K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[25] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Technical manual and affective ratings. Technical report, University of Florida, Center for Research in Psychophysiology, Gainesville, USA., 1999.

[26] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993.

[27] E. Marosi, O. Bazán, G. Yañez, J. Bernal, T. Fernández, M. Rodríguez, J. Silva, and A. Reyes. Narrow-band spectral measurements of eeg during emotional tasks. *The International Journal of Neuroscience*, 112(7):871–891, 2002.

[28] M. M. Müller, A. Keil, T. Gruber, and T. Elbert. Processing of affective pictures modulates right-hemispheric gamma band eeg activity. *Clinical Neurophysiology*, 110(11):1913–1920, 1999.

[29] R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, MA, USA, 1997.

[30] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[31] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61(1):5–18, 2006.

[32] J. A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[33] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch. Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007.

[34] D. Sander, D. Grandjean, and K. R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005.

[35] A. Schlögl and C. Brunner. Biosig: A free and open source software library for bci research. *Computer*, 41(10):44–50, 2008.

[36] R. Sinha, W. R. Lovallo, and O. A. Parsons. Cardiovascular differentiation of emotions. *Psychosomatic Medicine*, 54(4):422–435, 1992.

[37] R. M. Stern, W. J. Ray, and K. S. Quigley. *Psychophysiological Recording*. Oxford University Press, Inc., 2 edition, 2001.

[38] E. van den Broek, J. H. Janssen, J. Westerink, and J. A. Healey. Prerequisites for affective signal processing (asp). In P. Encarnao and A. Veloso, editors, *BIOSIGNALS*, pages 426–433. INSTICC Press, 2009.

[39] E. van den Broek, M. Schut, J. Westerink, J. van Herk, and K. Tuinenbreijer. Computing emotion awareness through facial electromyography. *Computer Vision in Human-Computer Interaction*, pages 52–63, 2006.

[40] C. M. van Reekum, T. Johnstone, R. Banse, A. Etter, T. Wehrle, and K. R. Scherer. Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, 18(5):663–688, 2004.

[41] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133, New York, NY, USA, 2007. ACM.

Detecting affective covert user states with passive Brain-Computer Interfaces

Thorsten O. Zander
Team PhyPA
TU Berlin, Germany
tza@mms.tu-berlin.de

Sabine Jatzev
Team PhyPA
TU Berlin, Germany
sja@mms.tu-berlin.de

Abstract

Brain-Computer Interfaces (BCIs) provide insight into ongoing cognitive and affective processes and are commonly used for direct control of human-machine systems [16]. Recently, a different type of BCI has emerged [4, 17], which instead focuses solely on the non-intrusive recognition of mental state elicited by a given primary human-machine interaction. These so-called passive BCIs (pBCIs) do, by their nature, not disturb the primary interaction, and thus allow for enhancement of human-machine systems with relatively low usage cost [12, 18], especially in conjunction with gel-free sensors. Here, we apply pBCIs to detect cognitive processes containing covert user states, which are difficult to access with conventional exogenous measures. We present two variants of a task inspired by an erroneously adapting human-machine system, a scenario important in automated adaptation. In this context, we derive two related, yet complementary, applications of pBCIs. First, we show that pBCIs are capable of detecting a covert user state related to the perception of loss of control over a system. The detection is realized by exploiting non-stationarities induced by the loss of control. Second, we show that pBCIs can be used to detect a covert user state directly correlated to the user's interpretation of erroneous actions of the machine. We then demonstrate the use of this information to enhance the interaction between the user and the machine, in an experiment outside the laboratory.

1. Introduction

The introduction of methods from statistical machine learning [1] to the field of brain-computer interfacing (BCI) had a deep impact on classification accuracy and it also minimized the effort needed to build up the skill of controlling a BCI system [2]. This enabled other fields to adapt methods from BCI research for their own purposes [18]. A particularly exciting development is the adoption of BCI technology into general Human-Machine Systems (HMS), i.e. for healthy users. In the context of HMS, a BCI constitutes a new communication channel, which enables direct insight into the cognitive and affective user states in response to the environment and technical system (see figure 1). However, not all types of BCIs are equally applicable in this context.

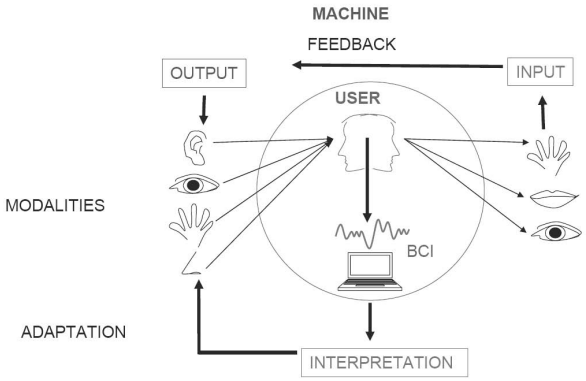


Figure 1. The feedback cycle of Human-Machine Interaction and its augmentation by BCI input.

We categorize the methods commonly applied in BCI research into active and reactive. By the term active BCI we denote BCIs which utilize brain activity of direct correlates of intended actions as input. This includes the detection of motor imagery or execution for active control, or the control over slow cortical potentials.
A reactive BCI is still controlled via intended actions. However, in contrast to the active BCI, features are not derived from direct correlates to these actions, but from cognitive reactions to exogenous stimuli, as e.g. in the P300 speller.
According to this line of thought, we now define passive BCI [4, 17]. Passive BCIs are not used with the purpose or ability of explicit voluntary control. Instead, they infer cognitive states which are already present within in the primary interaction in a human-machine system. Examples are brain states or cognitive events that are automatically and implicitly induced by the primary interaction. Hence, the inclusion of a passive BCI channel to an existing human-machine system does not directly interfere with the primary

978-1-4244-4799-2/09/$25.00

Type of BCI	Based on features from	Used for
Active	dedicated and intended action encoding a command focussed on directly generating a BCI detectable signal	direct control.
Reactive	dedicated perception on an exogenous stimulus focussed on indirectly generating a BCI detectable signal	direct control via external stimuli, brain switch
Passive	changes in cognitive states or occuring of cognitive events directly resulting from the current Human-Machine Interaction without the necessity of any additional user effort	implicit interaction, providing information on (affective) covert user states

Table 1. Categorization of BCI systems based on type of the used features and fields of application. The classical definitions of active and reactive BCIs are augmented by the definition of passive BCIs. These benefit from high classification rates and a low usage effort.

mode of interaction, and the passive BCI forms a secondary communication channel (see table 1).
EEG features based on implicit brain reactions to the environment also seem to be more robust in comparison to features utilized for active BCIs, which are often dependent on intended user actions and more variable in nature. This might be due to the fact that passive BCI features usually depend on automatic processes of cognition which are not as easily modulated by conscious processes. For these reasons, passive BCIs are readily applicable in the general Human-Machine Systems context. Especially, they allow for insights into user states, which are hard to infer from the users behaviour or other exogenous factors. We call these user states 'Covert User States' (CUS) analogous to the term covert attention (defined in [11]). CUS can refer to the user's interpretation of the current interaction states, which is usually not communicated directly to the technical system, but only conveyed indirectly by reactive user actions in response to this interpretation and task goal. This makes an interpretation and therefore an adequate adaptation for the machine to the user's need difficult. For instance, the user often expects a specific response of the machine to his behaviour. In adaptive systems this expectation is not always fulfilled, as the adaptation mechanism usually is based on a fixed rule system interpreting the user's behaviour. Hence, a corrective action by the user is necessary, which may disturb previous goals and strategies. The information of the user's actual interpretation of the situation, e.g. the CUS 'This is wrong!', could fundamentally augment the rule system of the machine and thereby enhance the adaptation performance.
In addition, CUS refer to cognitive and affective events or states that might be visible in explicit behaviour but are ambivalent until a direct categorization is possible. Since there is no direct communication between man and machine - with respect to the motivational and emotional response of the user - an adequate adaptation is difficult. Consequently, mental parameters related to these processes are an interesting addition to human-machine systems.
Affective CUS which are possibly detectable within the passive BCI framework range from mental workload, relaxation, surprise, and attention to arousal, frustration and more. Here, we are going to investigate more specific affective CUSs. In the subsequently presented scenarios, we investigate the impact on brain responses of misbehaviour of the machine, and their detectability via passive BCIs. Two different approaches are presented, one focusing on the latent state of loss of control over a system, and the other focusing on the immediate response following a faulty and surprising interaction state.

2. Methods

2.1. Specifications of our BCI system and experimental design

2.1.1 Recording

The EEG system has 32 channels of Ag/AgCl conventional (EasyCap) as well as impedance optimized (ActiCap) electrodes. Signals are amplified by a BrainAmp DC system and recorded by the BrainVision Recorder (BrainProducts). The electrodes are distributed on standard 10/20-based caps with 128 positions. Depending on the type of experiment, they are placed over according parts of the cortex. Additionally, we record electrooculogram (EOG) for controlling feedback-induced correlated eye movements, and electromyogram (EMG) on the relevant limbs, for protocolling correlated movements. Both are bipolarly multiplexed by a BrainAmp (ExG) system and derived with Ag/AgCl electrodes. In order to retain information on exogenous factors, we also record ambient temperature and noise level within the laboratory.

2.1.2 Experimental Conditions

The stimulus presentation in calibration phases before online feedback is designed for providing high control over exogenous and correlating factors besides the one of interest.

This control is relaxed in certain online feedback sessions to allow for a more realistic mode of interaction. A realistic HMS interaction mode is characterized by a distinctive motivational component with regard to the user, whose behaviour is driven by a certain task goal he likes to achieve. In this sense, the experimental paradigm can mimic real world scenarios, where mostly the motivational aspect is modulating the user's mental states and actions. This can be accomplished by putting the experimental paradigm into a game context, as is the case for the RLR-Game (see section 2.2). This decrement of control over factors might allow for a higher number of artifacts but does decrease the signal to noise ratio. Subjects have been introduced to the main factor of investigation by an instructor. Experimental tasks have been presented in a standardized way on the screen of the Feedback Unit. The course of the experiments contained several breaks for relaxation and recovering of the subjects. Subjects gave information on their overall state and their impressions on different blocks of the experiment by answering questionnaires. All subjects are from age 18 to 45 with German as primary language. The groups of subjects are of mixed and approximately balanced gender. Each subject was paid 20 Euros after completing the experiment.

2.1.3 Analyses

Classification: For offline analyses, all feature extraction methods, including filtering and resampling, are applied in a strictly causal way. Classifiers are chosen from several linear (LDA, rLDA, SVM) and non-linear (kernel SVM, RDA, GMM) methods. In all analyses presented subsequently, (regularized) LDA was the best performing classifier and was therefore selected. Classification accuracy was estimated by 10x(10[x5]) [nested] crossvalidation if not otherwise stated. Results from offline analysis are derived from strictly separated training and test blocks. Significance statements are substantiated by standard T-Tests and F-Tests without assumptions on the type of underlying distributions.

Feature extractors: For the extraction of features correlating to finger movements, two methods are used. First, the Common Spatial Patterns for Slow Cortical Potentials (CSPfSCP) algorithm [6]. CSPfSCP aims to find linear combinations (patterns) of EEG channels such that the detection of each trial projected according to these patterns is most discriminative (i.e., differs maximally between the two classes). This version is optimized to detect the deflection of the readiness potential (Bereitschaftspotential). This is an SCP indicated by contralateral low-frequency changes (1-5Hz), in this case localized over motor cortex. A slow negativity can be observed prior to a movement, and the relative strength of this negativity in the channels over the left versus right cortical hemisphere is typically used to infer the laterality of the upcoming movement. And second, for the extraction of spectral features correlating to event related desynchronisations (ERD) we used another version of CSP, Spectrally Weighted CSP (SpecCSP) [15]. SpecCSP iteratively alternates between optimizing spatial and the spectral criteria. This way, the algorithm calculates a set of custom spatial projection together with a set of custom frequency filters. These are generated for discriminating ERD by logarithmic bandpower. For the single trial detection of other event related potentials, in this case the EEG pattern correlating to error responses of the brain, features have been extracted by a derivate of the pattern matching algorithm [3]. It has been extended for detection of several extrema of SCPs within a given epoch. The data is resampled at 100 Hz, epoched relative to the event marker and a FFT band pass filter was applied using a frequency range of 0.1 - 15 Hz. Pattern matching reduces the dimensionality of the EEG data, by partitioning trials into n time windows according to the proposed ERP shape and calculating the mean of each time window and single trial. This results in an n-dimensional feature vector for each of m EEG channels. The EEG data is mapped onto an n*m dimensional feature space, containing the class-specific features of the EEG signal.

Dependent measures for statistical non-stationarities: For detecting non-stationarities in movement-related features, i.e. for executed button presses of the left and right hand, we implemented two methods. Both measures were calculated relative to the training data's distribution of the initial calibration measurement. In the first one, we explicitly applied a measure of statistical deviation to feature distributions. We measured the Kullback-Leibler divergence (KLD) of the feature distributions for direct observation of non-stationarities [14]. The second method aims at detecting non-stationarities implicitly, by measuring the performance of a movement classifier. We define pseudo online classification rates (POC) for this purpose. POC rates were calculated by offline analysis serving as estimation for online classification results. They were determined as following: A classifier was trained on the initial training block. Then, this classifier was applied to every key press. An average of approx. 100 gradual classifier outputs in a one-second window before each key press was averaged and taken as the classifier's decision for this key press. The sign of this decision value (by default, left keys, on average, were assigned -1, right keys +1) was remapped according to the key actually pressed, such that correct decisions were assigned positive values and wrong decisions were assigned negative values. The result is a real number for each key that was pressed by the subject. Therefore, positive values would indicate overall correct classifier decisions, while values close to zero or negative would indicate overall wrong decisions.

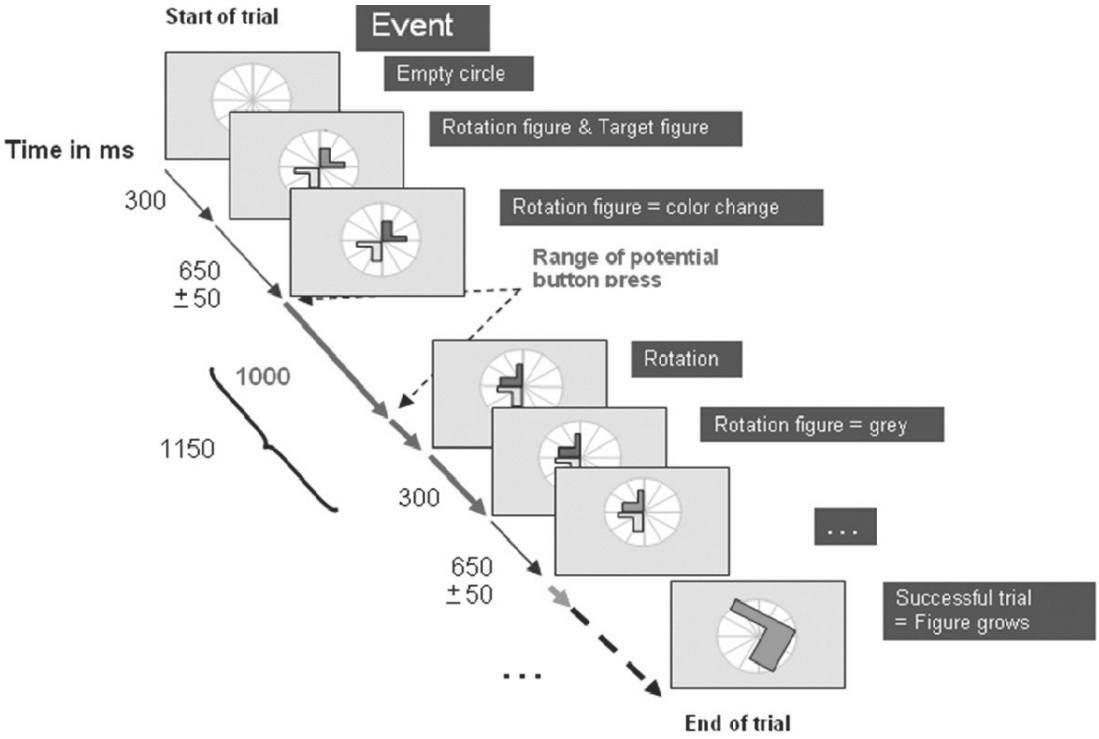


Figure 2. Example for single trial of the RLR-Design.

2.2. The RLR paradigm and its directed restriction, the RLR-Game

For both experiments, variants of the same experimental paradigm have been used, the Rotation-Left-Right paradigm (RLR) [9] and the RLR-Game. The RLR paradigm has been developed to mimic Human-Machine Interaction and to induce different mental user states by manipulating factors of interest (see figure 2). The goal of the experimental task is to rotate a stimulus clockwise or counter-clockwise (by a right or left key press, respectively) until it corresponds to a given target figure. The stimulus is either the letter "L" or "R", indicating the direction of rotation and left or right button press. While the colour of the stimulus is grey, it can not be rotated. However, every 1000 ms it changes into one of three colours, indicating A) the possibility to be rotated by a key press and B) the angle of rotation. If the stimulus lights up in red, the stimulus will rotate by 90 degrees, if it is yellow, by 60 degrees, and if it is green, by 30 degrees, upon key press. Each rotation has to be triggered, which only can be done once per colour change. The subject has to build up an efficient strategy for reaching the target: to rotate the starting stimulus as fast as possible on the target stimulus without rotating too far. The design can be played in two modes: The first was restricted to what we will call 'Full Control Mode' and the second, the 'Reduced Control Mode' included additional 'random states'. Random states are different from standard states in that they use a different rotation angle after key press, randomly selected from 90, 60, and 30 degrees. Consequently, in this case, the learned mapping rules do not apply anymore.

A derivate of the RLR paradigm is the RLR-Game, which is restricted to the colours green and red. It has also two stages, the "correct mode" and the "error mode" (see figure 3). In the "error mode", there will appear error states with a chance of 30%. The new angles in the error states are chosen to be always smaller than the ones from the corresponding standard case. Consequently, the colour red will result in 30° rotation (opposed to 90°), and the colour green will result in a 0° rotation (opposed to 30°). Secondly, the RLR-Game adds a second player, competing to the first one. Their performance is measured and fed back in form of points. A player gets a point when hitting the target earlier than his opponent. Hence the artificially induced machine error has a negative valence for the user, since it decreases his performance and might even lead to frustration.

3. Experimental Scenarios

In this section we are going to present two studies. In both the factor of investigation is the utilization of a specific CUS. The first study is based on the RLR design and handles the CUS of perceiving loss of control within human-

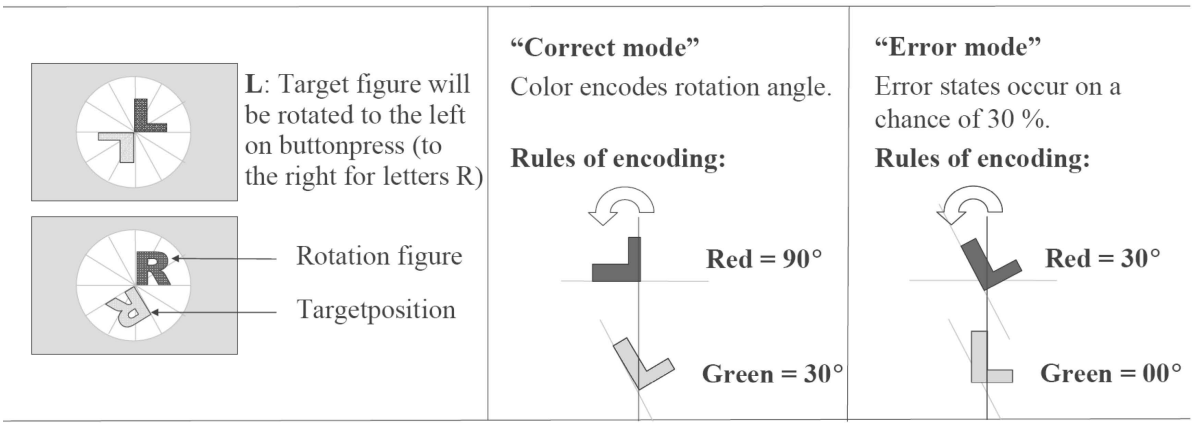


Figure 3. The RLR-Game and its two modes, correct and error. In error mode the system will react on a chance of 30% with a decreased rotation angle. A correct classification will then switch back to the standard rotation angle. On a false positive classification the angle would be correct and then, due to the classifiers mistake, be decreased.

machine interaction. In the second study we investigate the user's interpretation of automated adaptation within the RLR-Game.

3.1. Defining a pBCI for detecting the perceived loss of control within Human-Machine Interaction

3.1.1 Motivation

When conventional active BCI applications are transferred from the laboratory to interactive scenarios, influence of most interfering factors is lost, and such interference can lead to prolonged drops in performance. In this context, pBCIs may give insight into the underlying mental or affective user states, and related overall system states. In general, most HMSs are lacking information about the user's capability of handling the technical system, or whether the user is overwhelmed with the current task. Therefore, the machine is unable to adapt to the users needs and cannot supply the necessary support to avoid interaction mistakes.

3.1.2 Factor of investigation

An affective parameter that may underlie all of these situations is the loss of control (LoC), which makes conventional active BCI applications a good candidate for the investigation of the LoC. Moreover, a sufficiently universal method in the pBCI framework for detecting the LoC state may lend itself well to a much broader range of applications. This applies especially to those in which the primary interaction is performed by manual actions. As previously mentioned, accessing this state is difficult using conventional HMS channels, as LoC falls into the category of covert user states.

3.1.3 Approach

In this explorative scenario, we investigate the feasibility of designing a pBCI to detect the CUS of the loss of control over a task. It has been assumed that in the restricted context of active BCIs, e.g. control via imagined movements, the LoC manifests itself in non-stationarities in the underlying features. This leads to deviations from the feature statistics during the BCI calibration phase, with the consequence of degraded system performance [5, 14]. Theoretically, if present, this statistical behaviour can be detected passively, using the same features as a basis. However, to also allow for operation in more general HMS cases, features must be derived from executed movements, such as typing. Assuming that typing produces features that are similar to those occurring during imagined movements in conventional BCIs, we can reapply standard techniques for movement-related features to our new situation in a passive way. Standard feature extractors for imagined and executed movements are compared in their sensitivity with respect to the LoC, and thus in their predictive performance for use in a passive LoC detector.

3.1.4 Experimental Design

By utilizing the RLR Paradigm we have been able to artificially induce phases of reduced user control (phase BUc, see figure 5) by permuting the mapping between colours and angles of rotation. The learned rule system would not apply any more and therefore the user is confronted with an unexpected behaviour of the technical system, experiencing a loss of control of the task. In these experiments 22 subjects participated. We tracked features representing the primary mode of interaction, pressing a key, in the EEG data. Details on this study can be found in [9].

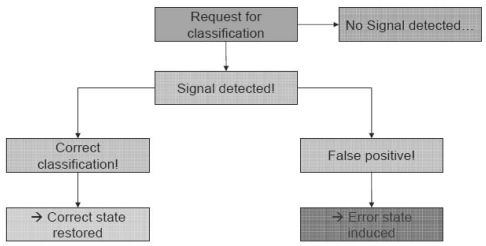


Figure 4. Scheme of the possible outcomes after the request for classification.

3.1.5 Features and analyses

Loss of control is accessed by analysing feature modulations of EEG patterns correlated to the executed hand movements. The EEG features that allow left and right hand movements to be discriminated fall into two categories: Slow Cortical Potentials (SCPs) and Event-Related Desynchronization (ERD) features. We have chosen features from both categories which have been extracted by Common Spatial Patterns for SCP (CSPfSCP) and Spectrally Weighted CSP (as described in section 2.0.3) for ERD from 200 ms of data prior to the button press. For the detection of LoC we have calculated the KLD on a moving window, containing the data of 10 button presses compared to the data from the initial training phase. Also, we estimated the POC for each buttonpress.

3.2. Applicability of a pBCI for enhancement of efficiency in HMS

3.2.1 Motivation

Errors in communication are highly relevant factors regarding the efficiency of HMS. Especially with regard to automated adaptation of the machine to the interaction mode of the user [10]. The machine tries to adapt to the behaviour and needs of the user. The currently used approaches are based on inferences of the user's actions or machine inputs. But a precise adaptation on this restricted information is hard to accomplish, because the mental states of interest are mostly CUSs. A wrong automation decision induces effects of surprise and frustration and in this respect, adaptation reduces the performance and the safety in HMS [13]. Additionally it triggers a correction action which enforces a shift in the intention focus of the user. According to this it reduces the overall acceptance of the adaptation and of the whole system. The goal of this scenario was to investigate a pBCI that is capable of communicating these error-related brain responses of the user to the technical system.

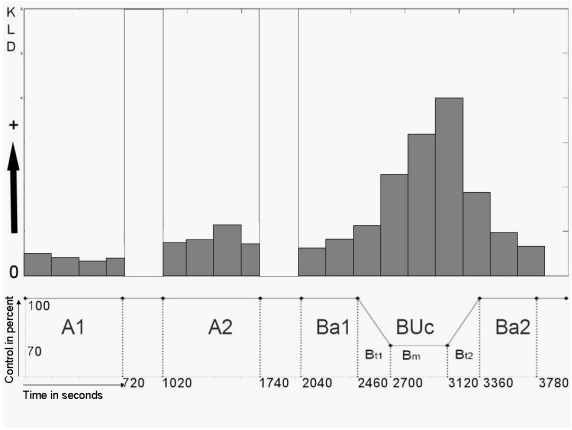


Figure 5. Grandaverage of the time course of the KLD for the CSP feature distribution through phases of full control (A1, A2, Ba1, Ba2) and phases of reduced control (BUc). In the first transition phase Bt1 the control is reduced to 70% gradually, it stays at this level in phase Bm and returns to full control in the second transition phase Bt2.

3.2.2 Factor of Investigation

The RLR-Game mimics the interaction in an HMS and allows for modelling an unexpected and negative effect, the error states. While this game is based on common interaction channels, we have added a secondary and passive BCI channel capable of automatically correcting the effects of reduced angles in the error states (see figure 3). This correction is triggered by an event-related potential reflecting the mental processing of an error trial. If it is correctly detected by the pBCI during an error trial, the rotation angle was set to the correct mapping. In case of a false positive the angle was reduced to that of a corresponding error state. Hence, each correct detection of an error brain response speeds the player up and a false detection slows him down. Therefore, if the classifier works properly, it will enhance the performance of the player and it will reduce it otherwise. See figure 4 for details.

3.2.3 Experimental Design

For keeping the environment as realistic as possible, we have chosen the Open House of the TU Berlin (LNdW 2007) as the setting. Four times two different players from the audience played the RLR game against each other (see figure 3) for a visualisation of the rule system). Each pair played three sessions of 50 trials. The first was for user training, without error states. In the second session we introduced the error trials. The automatic adaptation has been applied in the last session, only for one player.

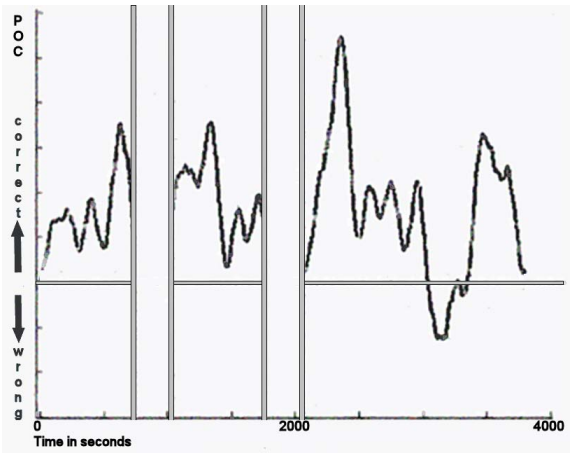


Figure 6. Grandaverage of the time course of the POC for the CSP features. A positive value indicates overall correct classification, a negative value reflects, that the classification accuracy is not reliable (< 50%). Strongly correlated to the decrease of the controlability of the system, the POC drops down.

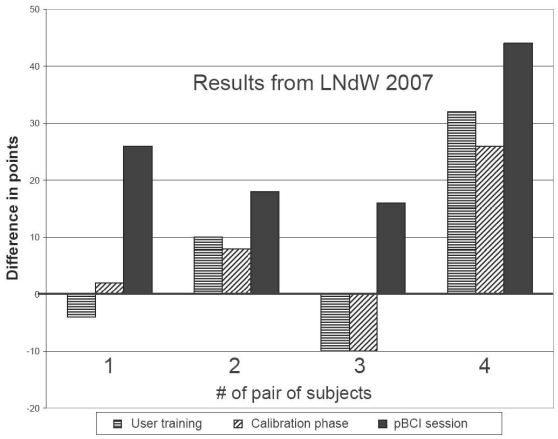


Figure 7. Results of the Open House of the Technical University of Berlin. The bars indicate difference in points of Player A (up) and Player B (down). Horizontal striped bars show the results from the phase "Subject Training", in which the game was set to Correct-Mode, allowing the user to learn the rule system. Results from the phase "Machine Training" are represented by the diagonal striped bars. Here, the system executed an error-state on a chance of 30%. On the resulting data a classifier was trained. In the third and last phase this classifier was applied to support Player A. As it can be seen in the orientation and magnitude of the black bars, the supporting system was successful.

3.2.4 Features and application

For detecting brain response relative to an erroneous rotation, we epoched the data from 0 to 800 ms relative to the stimulus rotation. On these epochs we applied the pattern matching method with 8 windows of 50 ms length starting at 300 ms after the event. This is resulting in a 256 dimensional feature space. Based on the features extracted from the data of the second RLR-Game session a classifier has been trained. In the third session this classifier was applied. On each trial a classification was requested directly after the stimulus rotation.

4. Results

The results of the LoC study (Figures 5 and 6) show that for the phases with full control (A1, A2, Ba1, Ba2) the variance of the averaged Kullback-Leibler divergence (KLD) is bounded for ERD features. In contrast, the phases of reduced control (BUc) reveal a significant ($p<0.05$) increase of the KLD, for ERD-based features. The POC drops down in phase Bm, which correlates significantly ($p<0.05$) to the course of the KLD. Hence, the KLD of this feature category is a measure strongly related to the perception of control by the user. Contrary there are no significant changes in the features extracted for slow cortical potentials.
Figure 7 shows the results from the sessions from the open house of the TU Berlin 2007, investigating the pBCI based on error potentials. During the third session one player was supported by the pBCI, correcting erroneous states of the machine by detecting brain activity induced by the respective machine error states. While the points have been equally distributed between session 1 and 2, the performance of all pBCI supported players has been increased significantly. The classifier hat an accuracy of 81.2% with error ratios equally distributed over the two classes.

5. Discussion

The results of the LoC scenario show that we have found a possible BCI measure for an affective mental parameter, which is sensitive with respect to loss of control in an executed movement task, relying on oscillatory features. Features based on the readiness potential, however, show no sensitivity. This indicates that the corresponding mental parameter could be detected passively, using SpecCSP and KLD as building blocks, and a technical system could be supported by it. As the loss of control is an important CUS to be transferred to the technical system in order to enable an adaptation of the system to the user's needs, this is an important route for further investigation. Especially, the online applicability and specificity of the inference have to be further investigated.
Erroneous system behaviour results in frustration of the user and a deteriorated man-machine interaction. The investigated pBCI online detection of brain responses to machine errors clearly allows for an enhancement of the human-machine interaction. Currently, a further study is being undertaken to validate the pBCI based on error responses, by investigating EEG patterns correlating to different error cat-

egories that induce similar EEG signals. Especially the idea of utilizing the EEG signal for sensing the subjective interpretation of current interaction states within HMS seems to be promising.

6. Conclusion

Here we gave examples of two types of pBCIs. One establishing an information flow from the human brain to the HMS, reflecting CUS correlated to current modes of interaction. The other one extracting the actual interpretation of dedicated system states from the users cognition. Both can be applied in the context of BCI for enhancing classification accuracy. First, for automated adaptation of BCI classifiers, and second, for correcting errors in Human-Machine Interaction as proposed in [3, 7]. In the more general context, our results show that pBCIs are suitable for an application in the field of HMS, providing information about the mental user state, which can only hardly be inferred by typical information channels in HMS. Our experiences with pBCIs show that these enable new channels of information within the interaction between man and machine. Next to an increased efficiency of work, automation technology has caused additional difficulties in HMS. This leads to errors and safety risks, mainly due to maladapted man-machine communication. pBCIs enable a direct access into CUSs, which is not currently accomplished by any other HMS method. Anyhow this is an important precondition for an optimal adaptation of automated agents, to make the man-machine interaction more efficient and less prone to errors. Here we were able to show that pBCIs are capable of detecting brain activity in response to machine errors and thereby enhancing automation adaptation. Additionally it seems to be fruitful to exchange experiences between the fields of HMS and BCI research, which will hopefully be done extensively in the near future. These studies could be a starting point for a whole series of new approaches. Currently we are investigating pBCIs for detection of mental workload, cognitive interpretation of the perception of human movements [8] and information on driver intentions. Please see www.phypa.org for details.

References

[1] B. Blankertz, G. Curio, and K. R. Mueller. Classifying single trial EEG: towards brain computer interfacing. In *Advances in neural information processing systems: proceedings of the 2002 conference*, page 157, 2002.

[2] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Mueller, and G. Curio. The non-invasive berlin braincomputer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539550, 2007.

[3] B. Blankertz, G. Dornhege, C. Schafer, R. Krepki, J. Kohlmorgen, K. R. Mueller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Trans Neural Sys Rehab Eng*, 11(2):127–131, 2003.

[4] E. Cutrell and D. Tan. Bci for passive input in hci. In *Workshop on Brain-Computer Interfaces for HCI and Games, CHI Conference*, 2008.

[5] G. Dornhege. *Increasing information transfer rates for brain-computer interfacing*. PhD thesis, Fraunhofer FIRST, IDA, 2006.

[6] G. Dornhege, B. Blankertz, and G. Curio. Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials. In *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*, page 595598, 2003.

[7] P. Ferrez. *Error-related EEG potentials in brain-computer interfaces*. PhD thesis, Ecole Polytechnique federale de Lausanne, 2007.

[8] M. Gaertner, T. Klister, and O. Z. Thorsten. Classifying the observation of feasible and unfeasible human motion. In *Proceedings of the 4th Int. BCI Workshop & Training Course*, volume 4, Graz, Austria, 2008. Graz University of Technology Publishing House.

[9] S. Jatzev, M. DeFilippis, C. Kothe, S. Welke, and M. Roetting. Examining causes for non-stationarities: The loss of controllabilityi s a factor which induces nonstationarities. In *Proceedings of the 4th Int. BCI Workshop & Training Course*, volume 4, Graz, Austria, 2008. Graz University of Technology Publishing House.

[10] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 30(3):286297, 2000.

[11] M. I. Posner and Y. Cohen. Components of visual orienting. *Attention and performance X*, page 531556, 1984.

[12] M. Roetting, T. O. Zander, S. Troesterer, and J. Dzaack. *Implicit interaction in multimodal human-machine systems*. Methods and Tools of Industrial Engineering and Ergonomics. Springer, Berlin, 2009.

[13] N. B. Sarter, D. D. Woods, and C. E. Billings. Automation surprises. *Handbook of human factors and ergonomics*, 2:19261943, 1997.

[14] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K. R. Mueller. Towards adaptive classification for BCI. *J. Neural Eng*, 3(R13-R23):11, 2006.

[15] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K. R. Mueller. Spectrally weighted common spatial pattern algorithm for single trial eeg classification. *Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep*, 40, 2006.

[16] J. J. Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157180, 1973.

[17] T. O. Zander, C. Kothe, S. Jatzev, R. Dashuber, S. Welke, M. D. Filippis, and M. Roetting. Team PhyPA: developing applications for Brain-Computer interaction. In *Workshop on Brain-Computer Interfaces for HCI and Games, CHI Conference*, 2008.

[18] T. O. Zander, C. Kothe, S. Welke, and M. Roetting. Enhancing Human-Machine systems with secondary input from passive brain-computer interfaces. In *Proceedings of the 4th Int. BCI Workshop & Training Course*, volume 4, Graz, Austria, 2008. Graz University of Technology Publishing House.

Practical study on Real-time Hand Detection

Jorn Alexander Zondag
Technical University of Eindhoven
Eindhoven, The Netherlands
jornzondag@gmail.com

Tommaso Gritti, and Vincent Jeanne
Philips Research Laboratories
Eindhoven, The Netherlands
{tommaso.gritti, vincent.jeanne}@philips.com

Abstract

In this paper we describe algorithms and image features that can be used to construct a real-time hand detector. We present our findings using the Histogram of Oriented Gradients (HOG) features in combination with two variations of the AdaBoost algorithm. First, we compare stump and tree weak classifier. Next, we investigate the influence of a large training database. Furthermore, we compare the performance of HOG against the Haar-like features.

1. Introduction

At the moment the typical human interface to devices or computers are keyboards, mice or remote controls. If we wish to communicate in a more natural way and possibly give orders to smart devices and electronics, hand gestures might be used. In order for the device to understand , an automatic gesture recognition method is needed.

The recognition task is typically preceded by hand detection. This is similar to the combination of the already widely used face detection and subsequent face recognition. Face detection and recognition are far from trivial and hands are even more challenging. This is due to the variability of the possible hand gestures. Hands are complex, deformable objects that are very dif cult to detect in dynamic environments with cluttered backgrounds and variable illumination. Several systems for gesture recognition as well as hand detection have been proposed [1, 2, 4, 5, 6, 9, 11].

Viola and Jones [13] proposed the combination of Haar-like features and boosted classi ers using the (cascaded) AdaBoost machine learning algorithm to train a detector for real-time face detection. The success of this combination for face detection has inspired researchers to employ this particular feature and machine learning algorithm for real-time hand detection as well [2, 5, 6, 9]. Consequent gesture recognition might then be performed using Hidden Markov Models [4] to link the transitions between different poses to form a gesture. The system proposed in [6] shows promising results, and uses face detection to provide a Region Of Interest (ROI) for initializing skin color segmentation. In [9] new speci c Haar-like features for hand detection have been proposed. The detector is trained using still images, which contain centered hands, with well-de ned gestures. In [5] an extension of [9] has been proposed, where the gestures are recognized using scale-space derived features. The reported experiments were carried out in a dynamic environment. In [2] the Haar-like feature set is extended with 45° tilted rectangular features. The recognition of different gestures is performed using several single gesture detectors in parallel. The system was tested under laboratory conditions, and its performance in dynamic environments is highly uncertain. In [11] detectors are constructed using Real and Discrete AdaBoost on four different databases. One database contains faces, the other three each contain a different hand pose. Haar-like features are combined in a single weak classi er to increase the discriminative strength. By evaluating the co-occurrence of features, more relevant spatial structure relationships can be encoded. The authors report a high accuracy on their databases when exploiting the co-occurrence of three Haar-like features in their weak classi ers. In [1] a system is proposed to nd both the hand and arm positions in sign language video sequences. Next to other cues, HOG image feature templates are constructed for each limb and used in a model to match limb con gurations to the unknown con gurations in the sequences. The reported accuracy is very high, however the processing time per frame is prohibitive.

In this paper, we focus on the task of real-time hand detection, without contextual information, in indoor environments with cluttered backgrounds and variable illumination, for a target application requiring low false positive rates. The detection task is characterized by a lack of consistent internal contrast in the hand combined with the complex background. We expect that a feature based on gradients would be able to encode the relevant structures. We verify this by adopting the HOG image features. Two variations of AdaBoost are tested, together with stump and tree weak classi ers. As a reference, we also analyze the performance of the commonly adopted Haar-like features. We

978-1-4244-4799-2/09/$25.00

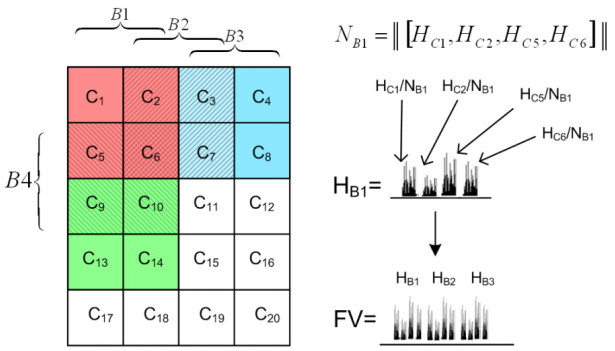


Figure 1. Overview of the HOG feature vector construction

show that the HOG features detectors can be trained with much larger databases than Haar-like features detectors, and achieve similar or better detection performance.

The remainder of this paper is structured as follows: Sections 2 and 3 describe image features and boosting algorithms used in our experiments, respectively. Section 4 discusses the acquisition of the hand database and the setup of the experiments. Section 5 shows the results of the experiments and Section 6 presents our conclusions.

2. Features

In the following sections, we will describe the HOG and Haar-like image features, which were successfully used for object detection in [3] and [13], respectively. We purposely discard color information for two main reasons: skin color is extremely unreliable in uncontrolled environments and it cannot be use in darker conditions, where active Infra-Red illumination is required.

2.1. Histogram of Oriented Gradients features

The Histogram of Oriented Gradients (HOG) [3] encodes the spatial distribution of local intensity gradients. A hand might be well detectable in a cluttered background by a characteristic local distribution of edges or intensity gradients. The HOG features are computed by dividing an image into small spatial regions called cells. For each cell a local 1-D histogram of gradient directions is accumulated over the pixels of the cell. The concatenated histogram entries of the cells form the HOG feature vector representation.

The gradient directions in the input image are computed, using discrete derivative masks like Sobel masks. Each cell-level histogram divides the gradient angle range into a fixed number of predetermined orientation bins. Each pixel in the cell votes (weighted) for an edge orientation, based on the orientation of the gradient element centered on it, into the orientation bins (angle ranges) of the cell's histogram. The orientation bins are evenly spaced over $0° - 180°$ or $0° - 360°$. In practice for invariance to illumination changes, a number of cells are combined to form a block and the cells in each block are contrast-normalized by using an accumulated measure of local histogram energy in the block. The blocks can be rectangular or circular log-polar.

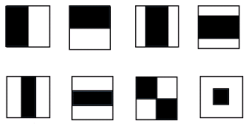


Figure 2. Set of Haar-like features

In Figure 1 an example overview shows how the HOG feature vector is constructed from a sample image. Each block histogram (B) is a concatenation of the normalized cell histograms (H_C/N), where both are 1-D vectors. The entire feature vector is then a concatenation of all block histograms.

2.2. Haar-like features

The Haar-like type of features encodes the oriented contrasts between regions in an image. For efficient implementation these features are usually designed as rectangular shapes with two or more non-overlapping black and white sub-rectangle areas and an optional rotation. The value of the feature is determined by computing the difference between the sums of pixels in the black and white area(s);

$$featurevalue = \sum_{area,white} i(x,y) - \sum_{area,black} i(x,y) \quad (1)$$

where $i(x,y)$ represents image intensity (pixel) value at position (x,y) in the image. Upright rectangular features can be computed efficiently and fast using an intermediate representation of an image, called integral image. This makes Haar-like features very suitable for usage in systems requiring high frame rates or high performance. In practice contrast stretching / normalization is used to improve robustness to differing lighting conditions. The feature vector is constructed from an example image by calculating a set of Haar-like features in all possible locations and scales and concatenating the (normalized) feature values into a 1-D vector.

3. Machine learning algorithms

For classification, a machine learning algorithm requires an input database with N training examples, translated into labeled feature vectors using an image feature (such as the described HOG or Haar-like features). This yields $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \epsilon \Re^k$ and $y_i \epsilon \{-1, 1\}$ (or $y_i \epsilon \{0, 1\}$). x_i is a K-component vector (K equals the number of features per example). y_i is called the class label

and indicates if the feature vector x_i was constructed from a positive or negative training example. In our case, a positive training example contains a hand, while a negative example contains anything but a hand.

In the following sections, we describe two commonly used variations of the AdaBoost algorithm, Gentle [8] and Discrete AdaBoost [7], and the different weak classifiers we used in our experiments. Furthermore, we present the cascaded classifier construction approach for AdaBoost [13].

3.1. Discrete AdaBoost

The name Discrete AdaBoost is adopted from the discrete classification made by the weak classifier. We define the discrete decision stump (weak classifier) $h_j(x)$ as;

$$h_j(x) = \begin{matrix} -1, & if \quad f_j(x) < \varphi_j \\ 1, & otherwise \end{matrix} \tag{2}$$

Where f_j is a feature, φ_j a threshold, $h_j(x) \epsilon \{-1, 1\}$, also see Figure 3. The usefulness of this definition becomes apparent when we recall that every component of the x_i in the training examples represents a feature (f_j). The weak learning algorithm is designed to select the single feature, which best separates the positive and negative training examples. For each feature dimension the weak learning algorithm determines the optimal threshold classification function, such that the weight of misclassified examples is minimized. Therefore on each round of training, AdaBoost will select the feature and thus it's corresponding weak classifier, determined by the weak learning algorithm, with the smallest error in classifying the training examples. AdaBoost combines weak classifiers to produce a powerful committee of weak classifiers, called a strong classifier. These weak classifiers only need to be slightly better than chance ($> 50\%$)[7]. The strong classifier is a weighted combination of (different) weak classifiers (features) plus a threshold. The algorithm is presented in Table 1.

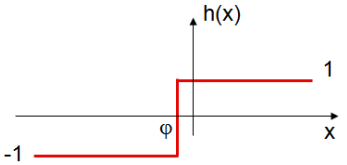


Figure 3. Decision stump for one feature (dimension). Samples below the threshold φ will be labeled -1 and above as 1

3.2. Gentle AdaBoost

The difference between Discrete and Gentle AdaBoost is found most notably in the employed reweighting of samples and their respective weak classifiers. The Gentle AdaBoost algorithm uses a different type of stump, called the regression stump. We define the regression stump (weak classifier) $g_j(x)$ as;

Input:
-Training examples $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \epsilon \Re^k$ and $y_i \epsilon \{-1, 1\}$ for negative and positive examples respectively
-Distribution D over the N examples
-Weak learning algorithm $weaklearn$
-Integer T specifying the number of iterations
Initialize weights $D_1(i) = \frac{1}{N}$
Do for $t = 1, \ldots, T$
1. Call $weaklearn$, which returns the weak classifier with $h_t : \chi \rightarrow \{0, 1\}$ from $F = \{h(x)\}$
with minimum error ϵ_t with respect to D_t,
$h_t = \arg\min_{h_j \epsilon F} \epsilon_j$, $\epsilon_j = \sum_{i=1}^{N} D_t(i)[y_i \neq h_j(x_i)]$
If $\epsilon_t > \frac{1}{2}$ stop
2. Set $\alpha = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$
3. Update the weights:
$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_t h_t(x_i)}}{Z_t}$
Where Z_t is a normalization factor chosen so that D_{t+1} is a distribution

The final strong classifier:
$H(x) = \begin{matrix} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq 0 \\ 0 & otherwise \end{matrix}$

Table 1. Discrete AdaBoost algorithm

$$g_j(x) = \begin{matrix} a, & if \quad f_j(x) < \varphi_j \\ b, & otherwise \end{matrix} \tag{3}$$

The output for this weak classifier is continuous, $g_j(x) \epsilon [-1, 1]$. a and b are determined by a weighted conditional expectancy on each side of the threshold φ_j. Therefore a and b will be different for each constructed weak classifier. If we take the regression stump in Fig. 4 as example, we can compute a, and b in a similar way by;

$$\begin{aligned} a &= E_D(y[x < \varphi]) \\ &= \frac{\sum_i^N D_i y_{i[x<\varphi]}}{\sum_i^N D_{i[x<\varphi]}} \\ &= \frac{\sum_i^N D_i \cdot 1_{[x<\varphi]} + \sum_i^N D_i \cdot -1_{[x<\varphi]}}{\sum_i^N D_{i[x<\varphi]}} \end{aligned} \tag{4}$$

Since the output of $g_j(x)$ is not on top of class labels, the error in classification made by the stump is calculated by weighted least squares and not by the misclassification error, as is the case for Discrete AdaBoost. The Gentle AdaBoost algorithm is presented in Table 2. Due to the output of the regression stump being bounded to $[-1, 1]$, the reweighting performed by this algorithm is more gentle than

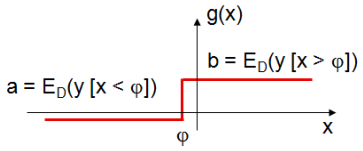


Figure 4. Regression stump for one feature (dimension). Samples below the threshold φ will have output level a and above b

Input:
-Training examples $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \epsilon \Re^k$ and $y_i \epsilon \{-1, 1\}$ for negative and positive examples respectively
-Distribution D over the N examples
-Weak learning algorithm $weaklearn$
-Integer T specifying the number of iterations
Initialize weights $D_1(i) = \frac{1}{N}$
Do for $t = 1, \ldots, T$
1. Call $weaklearn$, which ts the regression function $g_t(x)$ by weighted least squares of y_i to x_i with weights D_i, the optimized $g_t(x)$ is found through $\hat{g}_t = \arg\min_h E_D[(y - h(x))^2 | x]$
2. Update $H(x) \leftarrow H(x) + h_t(x) =$
$H(x) + \frac{E[e^{-yH(x)} y_i | x]}{E[e^{-yH(x)} | x]} = H(x) + E_D[y|x]$
3. Update the weights:
$D_{t+1}(i) = \frac{D_t(i) e^{-y_t h_t(x_i)}}{Z_t}$
Where Z_t is a normalization factor chosen so that D_{t+1} is a distribution

The nal strong classi er:
$H(x) = \begin{matrix} 1 & \sum_{t=1}^{T} g_t(x) \geq 0 \\ 0 & otherwise \end{matrix}$

Table 2. Gentle AdaBoost algorithm

is the case for Discrete AdaBoost, where the value of the α's (Step 2, Table 1) is unbounded and is even not de ned when the error is 0. Therefore Gentle AdaBoost is expected to be more robust to noisy data containing outliers, the weight of these examples can not increase as dramatically as is possible in Discrete AdaBoost [8].

3.3. Tree weak classifiers

Stump weak classi ers effectively split the training samples into two partitions using a threshold, labeling samples positive (object) or negative (non-object). Each partition may include a number of misclassi cations. With a tree [10] we can re ne the classi cation done by the initial stump. See Figure 5 for a visual representation, where N input training examples lie in a 2-D plane and a stump and a tree weak classi ers classify (divide) this same training space. Table 3 describes how we construct a tree weak classi er using stumps. When the tree weak classi er is used in combination with Discrete Adaboost we classify the training samples using a newly constructed tree and determine the corresponding α (Step 2. in Table 1) for the (tree) weak classi er in the same way as we would for a stump. However when using Gentle AdaBoost each leaf of the tree weak classi er is assigned an output value, determined by the weighted conditional expectancy of the samples that reach the leaf [8], see equation (4).

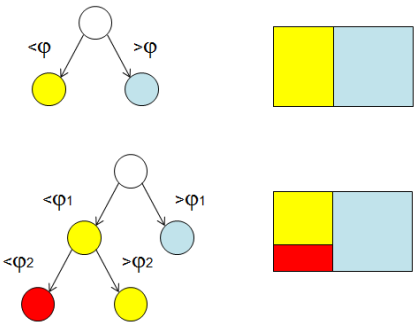


Figure 5. Stump and tree weak classi er dividing N examples

Input:
-Training examples $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \epsilon \Re^k$ and $y_i \epsilon \{-1, 1\}$ for negative and positive examples respectively
-Distribution D over the N examples
-Weak learning algorithm $weaklearn$
-Integer M specifying the maximum number of splits

Call $weaklearn$; select best stump on all training examples
Do for $m = 2, \ldots, M$
1. Call $weaklearn$ on all leafs; tting stumps on the training data which reaches that leaf
2. Add only one stump to the tree, select the stump which decreases the misclassi cation rate the most for its branch
3. Stop if the misclassi cation rate does not decrease for any branch

Table 3. Tree weak classi er construction

3.4. Cascaded boosting

To improve ef ciency, often a modi ed version of the AdaBoost algorithm is used, known as cascaded AdaBoost [13]. It is applicable to both Discrete and Gentle AdaBoost or any other variation of the algorithm. Instead of constructing one strong classi er to perform classi cation, a series or cascade of strong classi ers is created. These strong classi ers will have an increasing complexity (amount of weak classi ers). The main advantage is that early stages in the cascade (with low complexity) are able to label many of the

input samples already as non-object. These samples then no longer have to be evaluated by subsequent strong classi ers in the cascade, in contrast to non-cascaded strong classi ers where each sample passes through the entire strong classi- er before being labeled. Hence, there is a great reduction in the number of computations needed to process the input.

Performance goals in terms of detection rate and false positive rate (detection which is incorrect) for the cascaded classi er are de ned by respectively;

$$D = \prod_{i=1}^{L} d_i \tag{5}$$

$$F = \prod_{i=1}^{L} f_i \tag{6}$$

Where D is the detection rate of the cascaded classi er, L is the number of (strong) classi ers and d_i is the detection rate of the i-th classi er on the samples that get through to it. F is the false positive rate, and f_i is the false positive rate of the i-th classi er on the samples that get through to it.

Given the set of detection and performance goals, target rates can be determined for each cascade stage. The user selects the maximum acceptable rate for f_i and the minimum acceptable rate for d_i. Each stage of the cascade is trained by AdaBoost with the number of features (weak classi ers) being increased until the target detection and false positive rates are met for the current stage. The rates are determined by testing the current detector on a validation set of examples. If the overall target false positive rate is not yet met then another stage is added to the cascade.

4. Database & experiment setup

4.1. Database acquisition

To create our database we acquired clips using an HD video camera. We recorded four different persons performing an open hand pose on different backgrounds. The hand pose and the illumination conditions were varied during the recordings, as shown in Figure 7. The hand and ngers were moved unconstrained, but with the palm generally facing the camera. We also recorded clips of background , both indoor and outdoor. To create our positive examples we cropped the frames from the movies to the area around the hand, therefore there is only limited variation in scale.The negative examples were obtained by taking patches in random locations and different sizes in the frames of background movies.

Given this set of images, we constructed three databases. They contain the same examples in different sizes; 30x30 pixels, 60x60 pixels and 90x90 pixels. The databases contain over 140000 examples, the ratio between positive and negative examples is roughly $\frac{2}{5}$. Figure 6 shows the average image[12] of the positive examples in the database. It indicates there is a good amount of variation in the examples, because only a colored blob is visible, with nearly no visible remaining structure. As already mentioned, we discard color information.

4.2. Experiments

We separated $\frac{1}{3}$ of the databases for testing classi er performance (testing set). For training we used the remaining $\frac{2}{3}$ of the database examples, this is the total set of examples available for training (full training set). Using these sets we performed four experiments:

- Search of optimal HOG parameters. We trained detectors on half of the training set and veri ed performance on the testing set, for different HOG parameters.
- In uence of database size. We decrease the amount of training examples in the full training set in steps and train detectors for each training set size. The examples for each training subset are selected randomly from the full training set. To reduce the effect of random subsets, we repeat the training for a given set size multiple times and average the results.
- In uence of the database example size. Performance of the HOG features are compared when used for training with the three different image size databases.
- Comparison of HOG and Haar-like features. We train detectors using Haar-like features on the 30x30 example size database, for decreasing amount of training examples.

For all sets in the above experiments, we preserved the ratio between negative and positive examples present in the

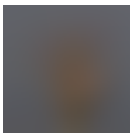


Figure 6. Average positive example image

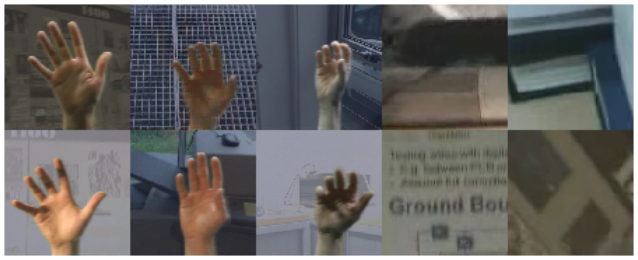


Figure 7. Database positive and negative examples

complete databases. The same examples were selected for each database, the only difference is in the example image size. Furthermore, we dedicated $\frac{1}{4}$ of training examples to validation, as required for cascaded boosting.

The results of the experiments can be found in Sections 5.2, 5.3 and 5.4 respectively.

5. Results

5.1. HOG parameters

In our experiments we have empirically determined the optimal parameters for the HOG features by performing an extensive parameter search on the databases. Performance for a combination of parameters was determined on the testing set. The most important parameters are the combination of block size, block stride and cell size and the number of angle bins and full angle range. These two combinations greatly determine the amount of detail that can be captured by the HOG feature and were the main focus of the parameter searches. To give an indication of the in uence of the parameters, a selection of results, from a part of the searches on the 30x30 pixel example size database, is presented in Tables 4, 5 and 6. The best performing parameters and the corresponding feature vector length are presented in Table 7. Shorter vector lengths are preferable as they can be calculated faster. The other optimal parameters which are not mentioned in the Table 7 are: we use, no normalization over blocks, the gradient magnitude for the weighted vote into the histogram bins and a diagonal lter kernel to compute the gradients.

Gradient computation	Performance
1D mask	0.9404
1D cubic corrected	0.9289
Diagonal lters	**0.9438**
Sobel lters	0.9422
Prewitt lters	0.9408

Table 4. The effect of different gradient computation on classi cation performance, for a xed set of parameters for block size, cell size, block stride, normalization type, voting method, angle range and number of bins, as shown in Table 7.

Voting method	Performance
Magnitude	**0.9438**
Magnitude Square	**0.9438**
Magnitude Square Root	0.9413
Binary voting	0.8930

Table 5. The effect of different voting methods on classi cation performance, for a xed set of parameters for block size, cell size, block stride, normalization type, gradient lter, angle range and number of bins, as shown in Table 7.

Cell Size (pixels)	2×2 cells	3×3 cells	5×5 cells	6×6 cells	10×10 cells
3×3	-	0.9170	-	-	-
6×6	0.9194	0.9296	-	0.9276	-
10×10	0.9249	-	0.9426	-	0.8732
15×15	-	0.9346	**0.9438**	-	-

Table 6. The effect of the cell and block sizes on classi cation performance, for the block stride equal to the cell size and a xed set of parameters for normalization type, voting method, gradient lter, angle range and number of bins, as shown in Table 7.

Database	bins	bl.sz.	cl.sz.	bl.str.	or.	length
30x30	9	15x15	5x5	5	180°	1296
60x60	9	20x20	10x10	10	180°	900
90x90	18	30x30	15x15	15	360°	1800

Table 7. HOG parameters per database; number of angle bins, block size (pixel), cell size (pixel), block stride (pixel), full angle range and the corresponding feature vector length

5.2. Database size

We investigated the in uence of the amount of examples available for training, on the detector performance. Figure 8 shows that the differences in average error rates between different combinations of AdaBoost and weak classi ers decreases when the amount of training examples increases (approaches the full database size), for the presented database. This behavior is seen for all three databases. Although the performance increase levels off for increasing database size, Figure 8 shows that more examples for training will increase the performance on the testing set. There is no great advantage of using one particular combination of weak classi er and boosting algorithm over another, if only the average testing error is considered. It also indicates that Discrete AdaBoost in combination with discrete decision stumps over ts on the training data, for small database sizes. There the average testing increases (detection rate and false positive rate both increase). This degradation in performance for small database sizes can be attributed to this algorithm's sensitivity to outlying or noisy data [8].

Database	Mean avg. testing error
30x30	2.88% ± 0.36%
60x60	2.19% ± 0.19%
90x90	**1.94% ± 0.20%**

Table 8. Mean average testing error of the different boosted weak classi ers on the full size databases

5.3. Example size

We compared the performance of detectors trained with the same examples, but of different image size. On the

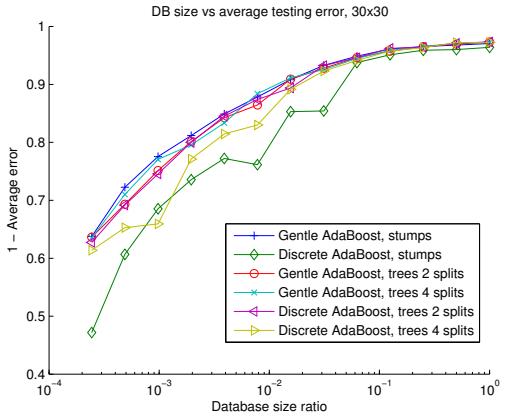


Figure 8. (1 - Average testing error) versus database size for different combinations of learning algorithms and weak classi ers, 30x30 pixel examples

whole, the lowest average error using HOG features is achieved on the database with the largest example size, see Table 8. However, the lowest average error for examples of 60x60 pixels is similar. In a practical implementation it would therefore make sense to use 60x60 pixel examples over 90x90 pixel examples. As this will allow the detection of smaller hands when using a sliding window method on multiple scales and faster computation of the feature vector, as the vector for 60x60 is half the length of the vector for 90x90 pixel examples (see Table 7).

5.4. Comparison between HOG and Haar-like features

To be able to judge our experimental results on, the detection rate, false positive rate and testing error, we determined classi er performance for classi ers using normalized Haar-like features on the 30x30 pixel example size database. The classi ers were trained with Discrete or Gentle AdaBoost and stump weak classi ers. We used the Haar-like feature set shown in Figure 2, computed for every possible scale and location, these features combine to a vector of 29415 elements. We use stump weak classi ers and Gentle AdaBoost. This combination consistently performs well and using a more complex tree weak classi er offers practically no gain in our case (see Figure 8).

Training the Haar-like feature detector is more demanding, since the Haar-like features have a feature vector length of over 30 times that of the optimal parameter HOG feature vector. During the experiments a great amount of extra memory was needed for the Haar-like feature training compared to the HOG feature. Therefore, only half the training set was used for training, in order to cope with the memory requirements and increased training time. We trained the HOG classi ers on half the database and used the same partitions of the training and testing sets and AdaBoost parameters to train the Haar classi ers, to ensure a fair evaluation.

Figures 9 and 10 show a comparison between the detectors using HOG features and Haar-like feature detectors on the 30x30 pixel example size database. The performance in terms of average testing error and average detection rate on the largest database size is roughly the same. However, Figures 10 and 9 show the HOG feature consistently achieving lower average false positive rates and lower average testing errors respectively for smaller training sets when combined with Gentle AdaBoost.

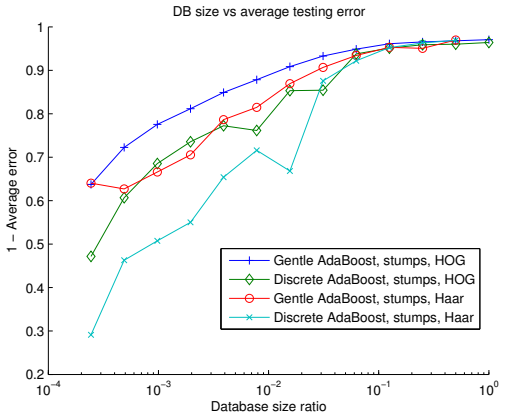


Figure 9. (1 - Average testing error), comparison between HOG and Haar-like image features on the 30x30 pixel example size database

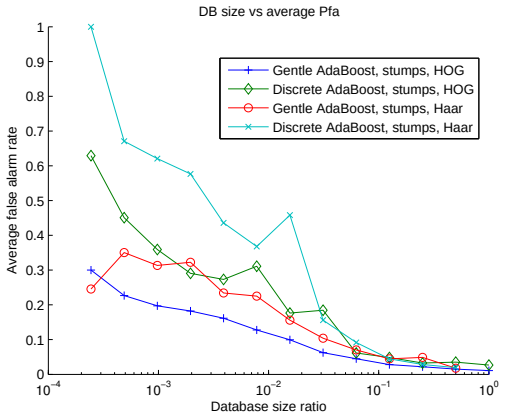


Figure 10. Average false positive rate, comparison between HOG and Haar-like image features on the 30x30 pixel example size database

In a hand detection system, Haar-like feature values can be determined in real-time by calculating one integral image per frame and a number of lookups in this image. For the HOG feature, when using integral histograms [14], as many integral images as there are angle bins need to be

calculated and histograms constructed for every (normalized) cell. HOG implementations can be sped up by fast approximations of the gradients, smart indexing into histograms in memory and using normalization factors from lookup tables. The short feature vector length of HOG features is an advantage. It reduces both memory requirements for training and training time. This is signi cant, since by expanding training databases with more examples, greater generalization can be achieved and detection accuracy improved. HOG features detectors can be trained with larger databases, due to the short vector length. The main disadvantage of the HOG feature is that optimal parameters need to be found for a given database.

We measured the detection speed, for an implementation of the hand detector using the scanning window method and, either the Haar-like, or HOG feature classi ers. We used integral images [13] for calculation of the Haar-like feature values and integral histograms [14] to calculate the HOG features, and no other optimizations. The measurements were performed for a 320 x 240 pixel video, on a Pentium 4. The Haar-like feature detector is around two times faster for this implementation, results are presented in Table

Classi er	Detection speed (FPS)
Gentle AdaBoost, Haar	**5.9**
Gentle AdaBoost, HOG	3.15

Table 9. Comparison of the detection speed (Frames Per Second) of the HOG and Haar-like feature classi ers.

6. Conclusions

We have presented an overview of the (cascaded) Gentle and Discrete AdaBoost algorithms, stump and tree weak classi ers and the HOG and Haar-like image features, which we used to construct a real-time hand detector. In the experiments, when presented with a large amount of training examples, AdaBoost with stump weak classi ers is able to construct a detector with similar performance as more complex tree weak classi ers are able to. The lowest average testing errors were achieved on the databases with large(r) example sizes, 60x60 and 90x90 pixel examples.

The experiments show a real-time hand detector using HOG image features can achieve similar performance to a detector using Haar-like features on the created databases, although Haar-like feature detectors can perform the detection roughly twice as fast. The HOG features have the advantage of having a much smaller feature vector than the Haar-like features, so HOG can be used in conjunction with much larger databases. The HOG features detectors consistently achieve better average false positive rates than Haar-like features detectors.

References

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British Machine Vision Conference*, 2008. 1

[2] Q. Chen, N. Georganas, and E. Petriu. Real-time vision-based hand gesture recognition using haar-like features. In *Proceedings Instrumentation and Measurement Technology Conference (IMTC)*, 2007. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886 893, 2005. 2

[4] N. Dang Binh, E. Shuichi, and T. Ejima. Real-time hand tracking and gesture recognition system. In *Proceedings Graphics, Vision and Image Processing (GVIP)*, 2005. 1

[5] Y. Fang, K. Wang, J. Cheng, and H. Lu. A real-time hand gesture recognition method. In *Proceedings IEEE International Conference on Multimedia and Expo (ICME)*, pages 995 998, 2007. 1

[6] H. Francke, J. R. del Solar, and R. Verschae. Real-time hand gesture detection and recognition using boosted classi ers and active learning. In *Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, volume 4872 of *Lecture Notes in Computer Science*, pages 533 547, 2007. 1

[7] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *In Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148 156, 1996. 3

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337 374, 2000. 3, 4, 6

[9] M. Kolsch and M. Turk. Robust hand detection. In *Proceedings International Conference on Automatic Face and Gesture Recognition*, pages 614 619, 2004. 1

[10] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classi ers for rapid object detection. In *DAGM-Symposium*, pages 297 304, 2003. 4

[11] T. Mita, T. Kaneko, B. Stenger, and O. Hori. Discriminative feature co-occurrence selection for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1257 1269, 2008. 1

[12] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. B. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition*, pages 29 48, 2006. 5

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 1, 2, 3, 4, 8

[14] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1491 1498, 2006. 7, 8

Personality Differences in the Multimodal Perception and Expression of Cultural Attitudes and Emotions

Céline Clavel
LIMSI-CNRS
BP133 – 91403 Orsay cedex - France
celine.clavel@limsi.fr

Albert Rilliard – LIMSI-CNRS
Takaaki Shochi – Kumamoto University
Jean-Claude Martin – LIMSI-CNRS
{rilliard,martin}@limsi.fr,
shochi38@gmail.com

Abstract

Individual differences have been reported in the literature on nonverbal communication. Recent development in the collection and evaluation of audiovisual databases of social behaviors brings new insight on these matters by exploring other types of social behaviors and other approaches to individual differences. This paper summarizes two experimental studies about personality differences in the audiovisual perception and expression of social affects. We conclude on the potential of such audiovisual database and experimental approaches for the design of personalized affective computing systems.

1. Introduction

Individual differences in non-verbal behaviors have already been reported in the literature (e.g. cultural differences with respect to rules for displaying emotions in public vs. private settings ; impact of introversion of the encoding on nonverbal behaviors) [1, 9].

Recent development in the collection and evaluation of multimodal databases of social behaviors [6, 10] enables to bring new insight on these matters by exploring other types of social behaviors and other approaches to individual differences.

This paper summarizes two experimental studies about personality differences [3] in the audiovisual perception and expression of culturally encoded social affects.

The study described in section 2 illustrates how the perception of audiovisual expressions of controlled attitudes depends on Japanese and French culture. We also introduce new analyses about the impact of personality traits on the perception of such behaviors.

The study described in section 3 considers another approach to personality (cognitive style) and studies its relation with multimodal expressions of emotions. We conclude on the potential of such database and experimental approaches for the design of personalized affective computing systems.

2. Study #1: Cultural and personality traits differences in social affects

Social affects, or attitudes, are expressions encoded in a language and a culture. As socially encoded tools, they are learned by children during developmental phase and have to be learned by foreign language students if such attitudes are not shared by the two languages [5]. They are thus quite relevant for the study of cultural differences in social behaviors.

2.1. Social affects in Japanese and French

Rilliard et al. [13] have described the production and the perception of respectively 12 and 6 audio-visual prosodic attitudes in Japanese and French, produced on an affectively neutral sentence. The 12 Japanese attitudes are: *doubt-incredulity* **(DO),** *obviousness* **(EV),** *surprise exclamation* **(SU),** *authority* **(AU),** *irritation* **(IR),** *arrogance* **(AR),** *sincerity-politeness* **(SIN),** *admiration* **(AD),** *kyoshuku* **(KYO),** *simple-politeness* **(PO),** *declaration* **(DC), and** *interrogation* **(IN). Three politeness expressions are presented:** *simple-politeness*, **the** *sincerity-politeness*, **used when the speaker is socially inferior to its interlocutor, and** *kyoshuku*, **a typically Japanese expression described by Sadanobu ([15], p. 34) as "***a mixture of suffering ashamedness and embarrassment, [which] comes from the speaker's consciousness of the fact that his/her utterance of request imposes a burden to the hearer***." The 6 French attitudes are:** *declaration* **(DC),** *interrogation* **(IN),** *obviousness* **(EV),** *surprise exclamation* **(SU),** *doubt-incredulity* **(DO),** *suspicious-irony* **(SC). Shochi [17] provides complete definitions for the attitudes.**

All attitudes, performed by two speakers for each language and filmed in a sound proof room, were then perceptually evaluated by native listeners of each language. Both the audio and visual modalities were presented alone to listeners, before an audio-visual presentation. For each utterance, listeners had to rate the perceived attitudes out of the 12 (or 6) possible attitudes. Recognition scores for each attitude, as well as confusions between attitudes were analyzed for each modality.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

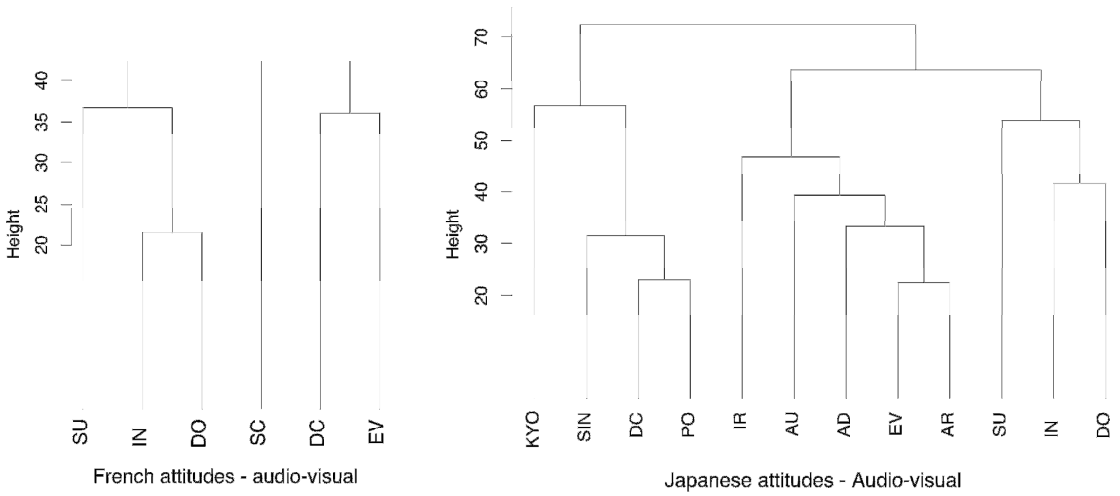


Figure 2: Hierarchical clustering of the 6 (for French, left) and 12 (for Japanese, right) attitudes, as perceived by native listeners in the audiovisual modality. Note that the Japanese admiration is not well recognized. Reproduced from [13].

2.2. Inter-speaker differences

The performances of the two speakers of each language have then been compared in order to rate the effect of the individual strategy and their ability to express attitudes. Their relative performances for each attitude are represented in figure 1.

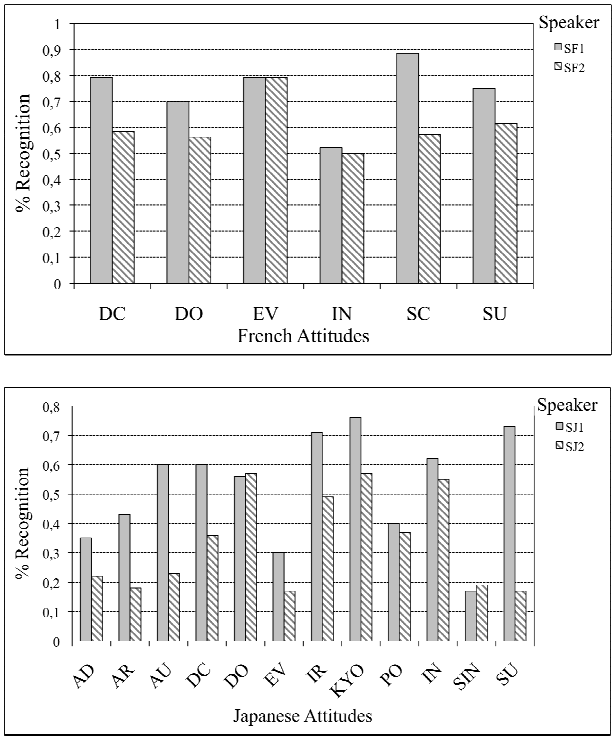


Figure 1: percentage of recognition obtained by each attitude, all modalities averaged, for each speaker of French (top) and Japanese (bottom).

In each language the first speaker (SF1 for French and SJ1 for Japanese) was a trained teacher, used to express himself in front of an audience, while the second was an untrained speaker (SF2 for French, SJ2 for Japanese). Results clearly show that listeners can recognize all attitudes with a higher (or similar) score from the trained speaker of both languages, and this is true for each modality. Such inter-individual differences in the performance are relevant for the study of social affects. As social affects are encoded in a language and a culture, they are shared by all speakers of the language. But the ability to express them outside of an ecological communication context seems highly linked to the speaker's communicative skills.

2.3. Perception of social affects

Due to the important differences between speakers, the perception results obtained from native listeners for both languages have been analyzed only from the trained speakers. The analysis was performed in order to visualize the main perceptive confusions and differences between each pair of attitudes, via a hierarchical clustering. The complete analysis of these perception results was presented by Rilliard et al. [13]; here is only a short summary of the main trends. Figure 2 presents the classification of the attitudes for Japanese and French sets of attitudes.

Both Japanese and French social affects are grouped together into three main clusters, that may be related to the assertive and dubitative speech acts described by Brandt [2], plus a dimension of dominance described by Shochi et al. [18] as "*express[ing] the imposition of the speaker's opinion*" on his interlocutor. The Assertive cluster groups the Declaration and Obviousness expressions for French, and for Japanese the Declaration

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

plus the three levels of Japanese Politeness (*simple-politeness*, *sincerity-politeness* and *kyoshuku*). The Dubitative cluster groups for French and for Japanese the expressions of Surprise, Doubt and Question. The Dominance expressions are represented for French by the *suspicious-irony* and in Japanese by *irritation*, *authority*, *obviousness* and *arrogance*. The Japanese expression of *admiration* was not recognized efficiently by listeners, but was mixed with dominance expressions.

2.4. The Five Factor Model

One of the definitions of Personality is "a set of organized, stable and individualized behaviors" ; the goal of personality research is to try to describe, to explain and to predict this set [12]. In psychology, three types of different approaches study personality. The lexical approach to personality is the most developed. This approach proposes to classify the terms of natural language that are used to describe and understand human qualities. It enables to define constructs that have a relative temporal stability, a good predictive value, that are applicable to different cultures and that are socially important. These constructs correspond to "personality traits". Rolland [14] defines traits as "coherent sets of cognitions, emotions and behaviors that demonstrate a temporal stability and cross situational consistency". Such traits result from inferences and not from a directly observable reality. Different models and psychometric tools based on the lexical approach have been developed: the Eysenck Personality Inventory, 16PF, and NEOPI R. This tool is currently the most used. It is based on the five factor model.

This model proposed by Costa and McCrae [4] describes personality with two levels. The facets propose a fine and accurate description of personality. A domain corresponds to a group of facets. The big five model identifies 5 basic dimensions through factorial analysis.

Neuroticism is defined as a system regulating avoidance behaviors. Its role is to preserve the organism of pain by anticipating and by activating monitoring behaviors. A subject with a high neuroticism score presents a very critical vision of himself. He also has the tendency to feel frequently and intensively a wide range of negative emotions. The 6 facets of neuroticism are: Anxiety, Angry, Hostility, Depression, Self-Consciousness, Impulsiveness, and Vulnerability.

Extraversion is characterized as a system of regulation of approach behaviors. A high score on this trait reveals a strong sensitivity to pleasant stimuli and a tendency to feel frequently and intensively positive emotions. The 6 facets of Extraversion are: Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, and Positive Emotions.

Openness to Experience results in broad and varied interests, a capacity to search for and to live new and unusual experiences. It is a system of regulation of reactions to novelty. The 6 facets of Openness to Experience are: Fantasy, Aesthetics, Feelings, Actions, Ideas, and Values.

Agreeableness refers to interactions with others and especially to the tone of relationship with others. The 6 facets of Agreeableness are: Trust, Straightforwardness, Altruism, Compliance, Modesty, and Tender-Mindedness.

Conscientiousness relates to motivation, organization and perseverance in the conducts oriented towards a goal. A high score corresponds to a person who tends to set long-term goals, to organize her action and accepts the constraints bound to the satisfaction differed of the needs and desires. The 6 facets of Conscientiousness are: Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation.

2.5. Relations between social affects and personality traits

We have already seen that the individuality of speaker has a main impact in the expression of affects. In order to study the influence of personality traits of listeners on their ability to recognize social affects, the listeners of the perception test of both Japanese and French social affect have been asked to fill in a questionnaire to rate their big 5 coefficients according to the Five Factor Model (FFM).

30 subjects for French and 46 subjects for Japanese have completed both the questionnaire and the perception test.

In order to analyze a possible link between personality traits and recognition scores, the subjects with the most extreme score for each personality traits have been selected. The mean value for each trait for all subjects was calculated. Subjects selected for a personality trait should have received a score with a distance from the mean superior or equal to two times the standard deviation. The comparison was made between the recognition scores for the four speakers of Japanese and French and each personality trait. The two speakers with the lowest performances are studied here because a possible relation between the personality of listeners and their ability to decode less prominent cues is interesting. For each of the four personality traits (Agreeableness, which is not supposed to be linked with perception of affects, was not studied here), an ANOVA was performed on the recognition scores for each language, with two following fixed factors: (1) the 12 or 6 attitudes and (2) the subject's ranking for the considered personality trait (i.e. on which end of the traits' scale is the subject – for example subjects with high or Low neuroticism). The attitudes have no main effect on the results. The detailed interaction between attitudes and personality traits are detailed hereafter, with their significance rated thanks to T-tests.

The comparison of recognition scores and personality profiles of the listeners for the 6 French attitudes shows only effects for both speakers with the suspicious irony

attitude. Best recognitions scores are received by SF1, and subject with a high Openness performed even better for this speaker ($T(8)=4.00$, $p<0.01$). The SF2 speaker, which achieve lower performances, tend to be perceived with a similar accuracy than SF1 by listeners with a high degree of Neuroticism ($T(11)=-2.05$, $p=0.065$). This personality trait is linked to the perception of negative emotions. For the other expressions, the effects of personality traits concerns only speaker SF2. This observation shows the importance of personality profile for the detection of the weakest cues: some of SF2's attitudes, that receive low recognitions scores, are perceived by listeners with a high level on one personality trait relevant to the considered expression. For the dubitative attitudes, the figure is as follow. For *surprise*, listeners with a high level of Openness (i.e. people who appreciate new experience) have higher recognition scores ($T(8)=3.46$, $p<0.01$). Conversely for *doubt*, people with a low level of Openness (i.e. socially conservative) performed better ($T(8)=-2.68$, $p<0.05$). This show the subtle differences that exist between the different attitudes regrouped into the three large clusters by the analysis of perception results: perception of *surprise* is linked with openness to novelty, whereas *doubt* is linked with a tendency to established things. The *obviousness* expressed by SF2 is better perceived by listeners with a high level of Conscientiousness ($T(10)$, $p<0.05$) (linked to the perception of rules). As we will see for Japanese subjects, it could be linked with the ability to perceived conventionalized expressions. The perception of the two other French attitudes (*declaration* and *interrogation*) does not change according to the personality traits.

Amongst the 12 Japanese attitudes, the expressions of *arrogance* and *irritation* (expression of imposition of the speaker) are both better recognized when listeners have a high level of Conscientiousness (significant effect for the not very expressive speaker SJ2 for AR – $T(16)=2.26$, $p<0.05$, and an increase for SJ1 for IR – $T(16)=1.84$, $p=0.08$). High Conscientiousness refers to a focus on rules and conventions. Amongst the dubitative expressions, both SJ2 *interrogation* ($T(16)=2.17$, $p<0.05$) and *surprise* (performed by the expressive speaker SJ1 – $T(16)=1.92$, $p=0.07$) are better perceived by subject with a low Neuroticism. Amongst the assertive expressions of Japanese, the *simple-politeness* for both speakers ($T(16)=2.21$, $p<0.05$) as well as the *sincerity-politeness* for SJ1 ($T(16)=2.77$, $p<0.05$), both perceptually very close, received higher scores for listeners with a low Openness (i.e. people sensitive to conventions). *Declaration* performed by SJ1 receives higher score from listeners with high Conscientiousness (sensitivity to rules – $T(16)=2.56$, $p<0.05$), whereas the recognition of the very typical expression of *kyoshuku* does not change with personality traits. Finally, *admiration*, misperceived and mixed up with expressions of imposition, received lowest scores when listeners have a high Neuroticism (sensitivity to negative emotions – $T(16)=2.47$, $p<0.05$).

The observations made on the Japanese social affects can be put together with the results presented by Shochi et al. [7]. They asked both adults and children to rate the degree of politeness of 5 Japanese social affects: *simple-politeness*, *sincerity-politeness*, *kyoshuku*, *declaration* and *arrogance* (an impolite expression). Their results show the complexity of politeness expression in Japanese, which children around 10 years old have difficulties to judge adequately. The high level of social encoding of such expression may explain why listeners more sensitive to rules or conventionalized situations are more efficient to perceive them, when their expression is not straightforward for adults (as it is the case for audio-visual *kyoshuku*).

Influence of listener's personality traits on their perception of social affect may take several forms, according to stimuli. Subject with a higher sensitivity to a given expression may be able to understand a stimuli performed by a less expressive speaker, whereas other listeners might not get the expression from such a stimuli. In other cases, only the most sensitive listeners might understand an expression, when it is difficult to induce this expression outside any communicative context. These results stress the importance of both the cultural and the individual factors in the expression and perception of expressive speech.

3. Study #2: Cognitive style differences in the multimodal expression of emotion

3.1. Cognitive style in Psychology

In Psychology, cognitive styles characterize the mental activity rather than its content. They describe one's cognitive functioning but also certain aspects of one's social behaviors [8]. They relate to characteristic ways to perceive, remember, think and solve problems [11, 19]. They are deduced from our stable individual differences in the way of organizing and of dealing with information.

One of the most studied cognitive styles in psychology is the field-dependency dimension (FID). This cognitive style relates to the usual and favorite way of perceiving the information. People that are independent from the field (FI) have an analytical vision; they transform the information at their disposal to organize it according to their own criteria. Their conducts are rather directed toward objects and they tend to take the lead in social interactions. In contrast, people that are depending on the field (FD) are more sensitive to the perceptive and conceptual organization of the information. They are very attentive to interpersonal relations and tend to ask for information from others.

We suppose that such properties of mental activity might participate in the multimodal expression of emotion. Indeed, one goal of multimodal expressions of

emotion is to inform others of the way we process the current situation.

In this study, we explain how we collected a TV series corpus which is relevant for the study of the emotional multimodal perception and of the cognitive style perception. Our hypothesis is that cognitive style can be perceived in the multimodal expression of emotions. For example “FI” people do not consider much the point of view of others and tend to dictate their opinion. Thus they might not try to control their anger and might adopt broader and quicker movements than “FD” people.

3.2. A TV series video corpus

To explore the multimodal emotional expressions related to each pole of the cognitive style (FID), we applied a corpus-based approach. We selected several TV series. Such acted data provide recurrent behaviors displayed by a variety of characters over time when faced with different emotional situations. They enable to consider the role of various situations in the emergence of the emotional process and are informative on the stability of the emotional expression in the course of time according to the stable personality of the characters.

We designed a questionnaire for assessing the various parameters of the FID (orientation of the behaviors, type of interaction and type of perception). 50 subjects had to estimate the cognitive style of seven television series characters using this questionnaire. This enabled us to select five female characters recognized by the subjects either as being strongly FI or either being strongly FD. Video samples featuring emotional behaviors of these characters were collected. Five characters were selected and five emotion families were considered (happiness, anger, surprise, fear and sadness). 100 sequences have been selected, for a total duration of 2568 seconds.

3.3. Studying relations between cognitive style and emotion

299 students in Social Sciences participated in this study. Each subject viewed a sequence of the corpus without any sound so that his attention would be concentrated primarily on the non-verbal expression and not influenced by the semantic meaning of the situation. Subjects could watch each clip as many times as they wanted. Subjects had to respond to a questionnaire evaluating different aspects of the situation and of the character. 50 video sequences have been proposed to the participants.

The questionnaire is composed of four parts, each of them offering a list of claims. For each of them, the subject must estimate the degree of agreement according to a five points Likert scale. The first part of the questionnaire concerns the evaluation of the emotional situation. It refers to the model proposed by Scherer [16] and to the action tendencies defined by Frijda [7]. The second section of the questionnaire relates to the assessment of the cognitive style of the character. Three items relate to the orientation of conduct (towards other people vs. toward objects), three others to the type of perception (analytic vs. holistic), and five items concern the type of interaction (collaborative vs. dominant). The third set of items concerns the multimodal expressiveness and aims to assess the temporality of the expression, the quality of movement and the facial expressions. Finally, the last part of the questionnaire tries to define the emotion expressed by the character. The subjects had to select one or several labels among Anger, Surprise, Sadness, Fear, Stress, Joy or Neutral.

From the assessment of emotional information realized, three clusters were made. The clustering is a statistical approach that aims at regrouping data in several homogeneous groups. Here, the three built clusters define three distinct ways to approach the situation according to the various components proposed by Scherer. Cluster #1 is about to situations eliciting joy. Cluster #2 relates to fear and stress. Cluster #3 is related to mixed situation involving both positive and negative emotions. Therefore, we have two contrasted clusters and an intermediate cluster. Thus, we have an evaluation of the perceived cognitive style of the characters (via answers to the first questionnaire to determine the cognitive style of the characters) and the emotional characteristics of the situations (clusters refer to a type of assessment of the emotional situation).

An analysis of variance was performed in order to evaluate the impact of the cognitive style and the emotional context (represented by clusters) on the perception of multimodal expression and cognitive assessment. Results showed that the perceived action tendencies, the facial expressions displayed by characters, and the perceived movement quality vary according to the cognitive style and the emotional context.

For example, some action tendencies features of the submissiveness and inattention were more attributed to the "FD" character that is in context of stress and less assigned this character in a context evaluated as joyful. This is especially true for the item "x wanted to cry" (F (2, 265) = 5.75, p = 0.004). The same results was also observed for the action tendencies relating to avoidance as "x desires to stay away" (F (2, 265) = 3.80, p = 0.023). Concerning action tendencies relating to behaviors of approach, we noted that subjects assign these action tendencies more when the context is joyful and it is even truer if the character is "FD". Figure 3 shows the perceived “x wanted to dance” action tendency according to the emotional context and the character's cognitive style (F (2, 266) = 4.20 p = 0.016).

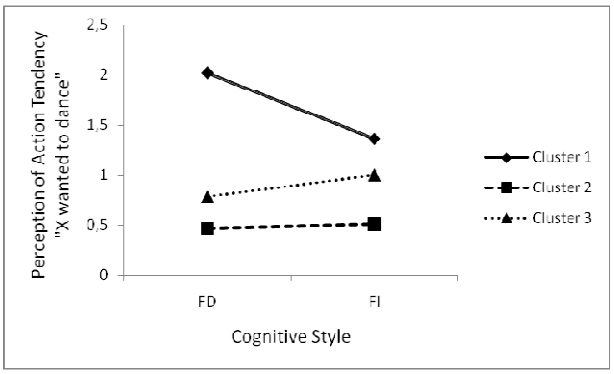


Figure 3: Level of perception of the action tendency according to cognitive style and type of cluster.

Such a video corpus of acted behaviors enables 1) to study non verbal communication in various situations, and 2) to explore the impact of personality on nonverbal behaviors. Thus, the perception of the multimodal emotional expression depends on the perceived emotional context and on the perceived cognitive style.

4. Conclusions

In this paper we described two experiments illustrating cultural and personality differences in the non-verbal perception and expression of two types of social affects (attitudes and emotions). These studies provide additional knowledge to the literature in nonverbal communication.

There is a great deal of variation in nonverbal communication between different social situations, and there are interactions between persons and situations [1].

Multimodal databases and experimental evaluations, as the ones described in this paper, enable to study these interaction effects and can be used to inform the design of personalized affective computing systems such as individual expressive virtual agents.

References

[1] Argyle, M.: Bodily communication (Second edition ed.). Routledge. Taylor & Francis., London and New York (2004).

[2] Brandt, P.A. (2008). Thinking and language. A view from cognitive semio-linguistics. In P. A Barbosa, S. Madureira & C. Reis (Eds.), *Proceedings of Speech Prosody* pp. 649-654. Campinhas, Brazil: Editora RG/CNPq.

[3] Clavel, C., Martin, J.-C. (2009). PERMUTATION: A Corpus-based Model of Personality and Multimodal Expression of Affects for Virtual Charcaters, *Proceedings of HCI International 2009* Town and Country Resort & Convention Center, San Diego: LNCS 5620.

[4] Costa, P.T., McCrae, R.R.: The NEO Personality Inventory manual. Psychological Assessment Resources, Odessa, FL (1985).

[5] Daneš, F.: Involvement with language and in language. Journal of Pragmatics, 22, 251-264. (1994)

[6] Douglas-Cowie, E., Cowie, R., Sneddon, C., C, Lowry, McRorie, Martin, J.-C., Devillers, L., Batliner, A. (2007). The HUMAINE Database: addressing the needs of the affective computing community. In A. Paiva, R. Prada & R. Picard (Eds.), *2nd International Conference on Affective Computing and Intelligent Interaction (ACII'2007)*, Vol. 4738, pp. 488-500. Lisbon, Portugal LNCS.

[7] Frijda, N., Kuipers, P., ter Schure, E.: Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212-228. (1989)

[8] Huteau, M.: Les conceptions cognitives de la personnalité. PUF, Paris (1985).

[9] Knapp, M.L., Hall, J.A. (Eds.): Nonverbal communication in human interaction (Sixth edition ed.). Thomson Wadsworth (2006).

[10] Martin, J.C., Paggio, P., Kuhnlein, P., Pianesi, F., Stiefelhagen, R.: Special issue on "Multimodal Corpora for Modeling Human Multimodal Behaviour". Journal on Language Resources and Evaluation, 41(3-4). (2007)

[11] Messick. The matter of style: manifestations of personality in cognition, learning, and teaching. Educational Psychologist, 29(3), 121-136 (1994)

[12] Mischel, W., Shoda, Y., Smith, R.E. (2004). Introduction to Personality: Toward an Integration, 7th ed. Hoboken, NJ: Wiley & Sons.

[13] Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., Aubergé, V.: Multimodal Indices To Japanese And French Prosodically Expressed Social Affects. Language and Speech, 52(2/3), 223-243. (2009)

[14] Rolland, J.P.: L'évaluation de la personnalité. Mardaga, Sprimont (2004).

[15] Sadanobu, T.: A natural history of Japanese pressed voice. Journal of the Phonetic Society of Japan, 8(1), 29-44. (2004)

[16] Scherer, K.R.: Emotion. In M.H.W. Stroebe (Ed.), Introduction to Social Psychology: A European perspective, Blackwell, Oxford (2000) 151-191.

[17] Shochi, T.: *Prosodie des affects socioculturels en japonais, français et anglais: à la recherche des vrais et faux-amis pour le parcours de l'apprenant.* Unpublished Ph.D. thesis, University Grenoble 3, France (2008).

[18] Shochi, T., Erickson, D., Rilliard, A., Aubergé, V., Martin, J.-C. (2008). Recognition of Japanese attitudes in Audio-Visual speech. In P.A. Barbosa, S. Madureira & C. Reis (Eds.), *Speech prosody 2008*, pp. 689-692. Campinas, Brazil: Editora RG/CNPq.

[19] Witkin, H.: A cognitive-style perspective on evaluation and guidance. Proceedings of the Invitational Conference on Testing Problems. (1973)

Social Signal Processing: What are the relevant variables? And in what ways do they relate?

Paul M. Brunet
Queen's University Belfast
Belfast, UK
p.brunet@qub.ac.uk

Gary McKeown
Queen's University Belfast
Belfast, UK
g.mckeown@qub.ac.uk

Roddy Cowie
Queen's University Belfast
Belfast, UK
r.cowie@qub.ac.uk

Hastings Donnan
Queen's University Belfast
Belfast, UK
h.donnan@qub.ac.uk

Ellen Douglas-Cowie
Queen's University Belfast
Belfast, UK
e.douglas-cowie@qub.ac.uk

Abstract

Studies of the processing of social signals and behaviour tend to focus intuitively on a few variables, without a framework to guide selection. Here, we attempt to provide a broad overview of the relevant variables, describing both signs and what they signify. Those are matched by systematic consideration of how the variables relate. Variables interact not only on an intrapersonal level but also on an interpersonal level. It is also recognised explicitly that a comprehensive framework needs to embrace the role of context and individual differences in personality and culture.

1. Introduction

A landmark review of psychology concluded that, after a century of research, it had reached the stage of trying systematically to identify the relevant variables [1]. It was not a nihilistic conclusion. The point was that a great deal of effort had been expended discovering that the relevant variables were not obvious.

Social Signal Processing should not need to spend a century reaching the same stage, because several disciplines – psychology being one – have worked to identify relevant variables. However, because the literature is large and diverse, it is easy to drift unintentionally into assuming that the variables are obvious. The aim of this paper is to offer the emerging computational discipline of Social Signal Processing a structured overview which helps to offset that tendency, and highlights some potentially relevant variables.

It is obviously not possible to review all that is known about social signal processing in a short paper. But it seems possible to provide a broad layout of the relevant literatures, and that is what the paper aims to do.

2. Alternative frameworks & models

Research from various disciplines has focused on determining how humans detect, interpret, and classify social signals, and consequently how this information affects behaviour during social interactions, and has proposed models and frameworks to explain and represent the processes involved in human-human interactions. These models are discipline-specific in the sense that they are produced by a methodology that lends itself to exploring some of the relevant variables and relationships involved in human social behaviour, but not all. Hence they tend to highlight one or two aspects of social behaviour and social cognition, and gloss over other relevant aspects. A key challenge for Social Signal Processing is to incorporate their different strengths into a comprehensive framework.

Experimental psychologists have worked extensively on the non-verbal signals that humans display and perceive signals during interactions, with particular emphasis on the face and the emotions that it conveys. Seminal work by Paul Ekman proposed 6 basic emotions (i.e. happiness, sadness, anger, fear, disgust, and surprise), each of which has a corresponding and universal facial expression [2, 3]. That has generated a large body of work on how humans can accurately detect emotional states by picking up on facial signals commonly associated with the basic emotions (e.g. a dropped jaw with relaxed lips and raised eyebrows associated with surprise [4, 5]. From there, research has diverged into topics from the ability of individuals with psychological disorders (e.g. autism [6], schizophrenia [7]) to identify emotional expressions, to sex differences [8], to developmental pathways [9]. Methods that promise objective measurement, from physiology to eye-tracking, have been eagerly embraced.

Linguists and psycholinguists, on the other hand, have focused on issues such as the structure implicit in dialogue. Some ideas, such as the pragmatic analyses associated with Grice and his colleagues, have become very well known. Others, particularly the analyses that deal explicitly with exchanges between two parties, are less familiar. For example, it is widely accepted that conversations occur on two tracks [10]. The first and main track represents the dialogue dedicated to the exchange of information. The second track represents the dialogue dedicated to the clarification and grounding of the main track's information.

Sociolinguists, in contrast, have focussed on the way speech encodes information about the social affiliations and aspirations of speakers, and relationships between

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

them. Features such as dialect and lexical selection play important, and quite complex roles here.

Anthropologists, like linguists, rely heavily on verbal data when developing models. They have proposed that there are universals in language that influence the foundation of social interactions. For example, an assumption of Brown and Levinson's model of politeness [11] is that all individual enter social interactions with the mutual understanding that both parties want to protect their 'face' and avoid harming the other person's 'face'. Face, the individual's public self-image which they do not want compromised by humiliation, has a positive and a negative subtype. Positive face represents the desire for approval by others, whereas negative face is the desire for freedom of action. All adult individuals know that they and everyone else has these desires.

One of the factors that distinguishes different approaches is that they are associated with different applications. From that point of view, Social Signal Processing is a discipline whose applications are very different from, for instance, language teaching, or psychological therapy. Hence it is appropriate that it should attempt to develop its own framework.

3. Towards a comprehensive framework

To achieve a framework that suits Social Signal Processing, a new model should strive to incorporate lessons from all the approaches outlined above. That cannot be done simply by adding together ideas from the various disciplines: the result would be an amorphous mass. To avoid that, a framework is needed that is capable of giving the relevant pieces a meaningful place. A first step towards that is to enumerate all the relevant elements – the potentially relevant signals, and the things they may signify; and to consider the ways in which the elements may relate.

3.1. Broad categories of variables

A social signal, as defined by Grammer et al.,[12] is a messenger carrying information between a sender and a recipient. The sender encodes information into the signal, which is then detected and decoded by the receiver. The signal is transferred over a communication channel which can take many forms. A description of the possible forms of signals follows.

3.1.1 Verbal characteristics.

Spoken language needs to be part of a comprehensive framework to understand social interactions. Not only is the content of the exchange important, individuals can also gather information based on the sentence structure, vocabulary, and purpose of the statements (e.g. self-disclosure, question, and request)[13, 14]. Additionally, many features of the way people speak are carry information not provided in the verbal content [14]. For instance, intonation plays an essential role in projecting and determining sarcasm or sincerity, nervousness or confidence, and approval or disgust, to name a few. The absence of that information has made computer-mediated communication more likely to lead to misunderstandings. When emailing and instant messaging first became popular, individuals did not know how to properly project or interpret tone (e.g. sarcasm) from the written exchange [16]. To help with that, emoticons and net lingo (e.g. writing 'lol' to signify 'laugh out loud' to project humorous tone) have arisen to inject tone into written exchanges.

3.1.2 Facial characteristics.

During social interaction, people extract a substantial amount of information about emotional states from facial expression rather than from verbal content [17, 18]. Recent research emphasises that in natural data, the signs are not simply archetypal 'snapshots': they are distributed over time [19] and linked to body movements [20].

It is also important not to focus exclusively on the facial surface. In particular, the eyes have been identified as a salient feature for conveying not only emotional states, but also determining intent [21]. Some functions are inherently interpersonal, such as establishing joint attention via eye gaze [22]. Joint attention refers to when two individuals are focused on the same event, object, or person. For example, if John makes a flattering comment about a third party who is standing near him to his friend Bill. Bill can identify who John is talking about by tracking John's eye gaze, and can then provide his own opinion. John does not have to state who he is talking about, his eyes gives that information. Channels of that kind are complex, but would pose real difficulties for an artificial system that could not use them.

3.1.3 Body characteristics.

In addition to the face, the body provides other relevant signals. Body posture (e.g. standing straight up, being slouched over, arms crossed) are useful social signals that have inspired the idea of body language [23]. Physical gestures also contribute to the information conveyed by the body. Many gestures can easily be identified and interpreted during social interactions (e.g. hand waving to say hello), which can indicate politeness, friendliness, aggression, and so on. Furthermore, the physical distance between both individuals can signal intent. For example, if a woman stands in very close proximity to a man, it could indicate attraction. If the woman keeps a noticeable distance between her and the man, it could suggest that she is not interested in his advances [24].

3.1.4 Physiological characteristics.

Physiological reactions can provide useful social and emotional information [25]. Some physiological reactions are undetectable without the aid of machines. For example, detecting someone's frontal EEG patterns is not possible unless they have electrodes on their head.

However, other physiological reactions are not only noticeable, but important social signals. For example, blushing, blinking rate, and sweating are signals that are detectable and may indicate being nervous, aroused, or embarrassed.

3.1.5 Other physical characteristics.

Other physical characteristics can influence social behaviour and signals, and provide useful information. The most salient of these observable physical characteristics related to gender. For example, short hair, a beard, an Adam's apple, and a flat chest are clear indications that the person is a man. These characteristics will influence the interpretation and the production of social behaviour. For example, a man is more likely to make emotional self-disclosures to a woman rather than another man [26].

Height and especially weight [27] are also social signals that can convey, but not always accurately so, useful information about the other person. A man who is physically fit is more likely to be asked questions about how to get involved with the local sports team than a man who is overweight.

Signals relating to age (apparent or chronological) also come into play. Apparent age is a social signal that the other person may detect and consequently affect their behaviour. Chronological age will influence the person's own behaviour and interpretation of the other person's signals. A 60 year old man and a 20 year old man may have different criteria for determining what signals of politeness.

3.1.6 High order characteristics

Information is also carried by what can be described as high order characteristics. These do not represent distinct variables, but instead refer to the quality of the other variable sets.

One such quality is the intensity of the signals and behaviour. The velocity and magnitude of a gesture, or the intensity of a facial expression, can be fundamental to the information that it provides.

Voluntary control over the signals can also be indicative of the person's intentions. A simple example is that of a smile. A child asks her father if he likes her art work. Smiling automatically would suggest that he genuinely liked the art; a forced smile could be read as concealing negative less positive opinion.

Patterning over time takes a wide range of forms. As a relatively simple example, recognising that an individual always speaks quickly or blushes easily and frequently affects their significance as social signals. At a much higher level, the pattern of interchanges between individuals in a group may indicate which has the role of chairperson, or acknowledged leader.

3.2. What variables may signify

Identifying the variables is one of two parallel tasks. The second is to establish what the variables may signify.

The most obvious is that the signals or behaviour are providing an explicit message that one person is trying to convey to the other. That is the natural way to think about linguistic communication, and it transfers to many kinds of non-linguistic communication too. However, that is only one possible kind of significance.

A second kind of communicative significance involves implicit meanings which the person may not consciously be trying to project. A clear example is that signs and behaviours can indicate the person's cognitive and affective states. For example, facial expressions are usually presumed to signify affective states. A smiling face is supposed to represent a happy affective state. Note, though, that there are other interpretations – a smile may be more analogous to a speech act.

A related set of possibilities, with strong links to philosophy and AI, is that variables may serve to convey a person's beliefs, desires, intentions, and attitudes. These may be conveyed in part or in whole by language, but many systems can contribute.

Various kinds of social position and relationship may also be signalled. Gender, age and fitness were mentioned in section 3.1.5. A wide variety of signals may be used to project dominance, authority, respect, or affinity. People's impressions of these attributes have a major impact on the form and success of social interactions with others.

More abstract analyses have been developed, and have a great deal to offer the field. There have been sophisticated attempts to operationalise the nature of a goal of a communication, invoking both proximal and distal explanations. Theories in ethology imply that there tend to be strong evolutionary pressures towards manipulation in social signals [28]. In contrast in theories concerning the evolution of human language there have been attempts to explain how cooperative goals can lead to cooperative social signals [29, 30]. More proximal intentions are perhaps easier to incorporate within a framework by acknowledging that signallers have goals even if it is as straightforward as a desire to sustain an interaction.

Behind each term in this section is a huge set of descriptive issues. For instance, affective states adequately is a research field in itself [30]. The options at this level are harder to articulate simply than the options for signals, and it is clear that Social Signal Processing needs systematic work on the problem.

3.3. Relationships among variables

3.3.1 Isolated & Combined Intrapersonal Effects.

A common paradigm in older research is to focus on a few selected variables and provide an explanation on how they convey information, how they are interpreted, and ultimately how they influence social behaviour. The resulting models focus on the isolated contributions of a variable. However, those isolated contributions are only simplified pieces of the overarching model. To compare the framework to the English language, each variable is like a letter. On its own, the contribution of

each letter (or variable) to the English language is minimal. The power of the system derives from the way letters can be combined to form words and sentences. Similarly, each variable (or category of variable) is important, but not to the exclusion of the others. To properly model how humans perceive and produce social behaviour, each of the variables must be accounted for, individually and interactively.

During communicative episodes, individuals do not isolate only one variable or category. Instead, information is gathered from all available social signals and is processed, analysed, and interpreted. For example, when having a conversation, nobody focuses only on the other person's facial expression because that information could be misleading. If the other person is smiling, it could indicate that they are happy. But if that smile is combined with blushing, stuttering, and gazing at their feet, the person is most likely not happy at all, but is trying to mask embarrassment. To properly understand the significance of the variables, a person cannot simply rely on another person's facial expression, but must also attend to all the relevant signals including the verbal information, the body posture, physiological reactions and so forth. All of these variables are individual and collective signals that humans attend to and process.

A comprehensive computation model would need to address the complexity of the relationship between the variables. Additionally, the variables do not all interact on the same level. Some variables function as a signal and others primarily function as behaviours. Signals and behaviours interact to influence the conceptual level which is the perception and interpretation of the signals and behaviours.

These integrative qualities of the variables have important practical ramifications. In a worst case scenario a signal may be embedded in a minimal increase in intensity of lots of variables. This means each of these variables would need to be assessed if a signal is to be detected. The reality for research projects with limited resources is a selection of a set of variables has to be made, usually guided by practical or historical and discipline related contingencies. As a consequence there may be certain signals that may not be detectable without a broad and inclusive set of variables.

3.3.2 Bidirectional Interpersonal Effects.

To complicate matters further, the bidirectional influence between the individuals engaged in the social interaction must be addressed by the model. While one person is detecting and interpreting the social signals conveyed by the other person, their own social signals are also being detected and interpreted. Social signals emitted are constantly being modified based on these interpretations. The social signals of Person A are influenced by the interpretations of the social signals of Person B, and vice versa [32]. Consequently, a comprehensive model must take into account all the potential social signals of Person A, their interactive relationships, the social signals of Person B, their interactive relationships, and the interactive relationships between Person A's and Person B's social signals. With the addition of further people to an interaction, greater sets of social signals and interactive relationships have to be accounted for by a model.

The interdependence of the Person A and Person B's signals and behaviours not only influence the conceptual mapping, but also the statistical analysis of the framework. The analysis of these interacting variables can be challenging with traditional experimental statistical techniques, and therefore is not appropriate for the analysis of dyadic communication. Instead, other types of statistics (e.g. multilevel modeling) must be applied [32].

Furthermore, a proper model needs to incorporate the ability to attend, perceive, and interpret multimodal signals. Communication is not solely auditory, nor is it solely visual; there are tactile and even olfactory elements. The signals available during communication cross over and interact between modalities. Consequently, the model must account for multimodal perception and attention. Fortunately, multimodal perception and attention is a growing area of research.

3.3.3 Logical relationships

It is tempting to assume that the logical nature of the relationships among variables is a simple conditional – if sign S, then condition C. However, much more complex types of relationship are common, if not the norm. Relationships often abductive (i.e. the best explanation for sign S is that condition C is present) or cancellable (i.e this is my inference about sign S, but new evidence may shows that it is wrong).

Sometimes the logical relationship can be reduced to a simple conditional by considering a complex of signals rather than one in isolation (not only a smile, but one with a particular time course, accompanied by movements of the head and shoulders). It is a key question how far that strategy can be taken.

3.3.4 Computing relationships.

Running through the discussion of relationships is a familiar computational issue. It is standard to contrast two approaches to constructing an internal representation of a relationship, conceptually-driven and statistically driven. Conceptually-driven models are based on theoretically proposed structures and relationships. These models have traditionally been fragile. Statistically-driven models involve models derived via machine learning by developing statistical relationships between the structures. These tend to be more robust, but at the expense of conceptual significance, and hence of generalisability. It is not obvious how Social Signal Processing should regard the two options. Some of the disciplines on which it draws are deeply sceptical of statistically-driven analyses, for non-trivial reasons. However, there are obvious practical reasons to use them in many applications.

3.4. Context of communication

3.4.1 Medium.

A framework that includes and models the relationships between all the possible variables needs to be flexible. Some modalities and corresponding variables may not be available during every communicative episode. The medium is a strong determinant of which modalities are present. In a face-to-face conversation, all modalities are usually available. The technologically-mediated means of communication occurs primarily in one modality at the expense of the others. For example, during telephone conversations, verbal exchange and voice tone are present. However, some modalities (e.g. visual system) are no longer providing social signals.

3.4.2 Setting.

The setting refers to the physical characteristics of the surrounding area. This includes the lighting, the space, and the scenery. Being in a darkly lit room results in a degradation of the social signals that are detectable, and this reduction in signal quality can consequently influence behaviour and the interpretation of the communicative episode.

3.4.3 Situation.

This relates to the purpose of the communicative episode instead of its setting. For example, social signals and behaviour will be different during a job interview compared to being in a cinema.

3.4.4 Person by Context Interactions.

Individual and cultural differences affect communicative episodes in two important ways. Firstly, an individual's ability to accurately detect and interpret social signals is influenced by their personality and their culture. For example, research has shown that individual differences in personality can influence children's accuracy in the categorisation of emotions [33]. Furthermore, individual and cultural differences will influence the social signals and behaviour emitted by a person. For example, an extremely shy person will produce signals of discomfort during social interactions, whereas his non-shy peer is less likely to do so [34].

The effects of individual differences are also highly dependent on the context. As proposed by Bem & Allen [35], some individuals are more strongly influenced by context, whereas others remain fairly consistent. Recent research in computer-mediated communication (CMC) has highlighted the person by context interaction by demonstrating that some individuals remain consistent in their behaviour across conditions with reduced social signals [36], whereas other can actually benefit from the reduction.

Another major factor is the familiarity and the nature of the relationship between two people. Fewer and subtler social signals can be detected by two people who have a close relationship [37]. For example, siblings can transmit more information with a look or one word than strangers can with an entire conversation.

3.5. Whose interpretation?

A final consideration when developing a computational analysis model is whose interpretation of the signals and behaviour the model is representing. The first consideration is whether the model should reflect how an observer would interpret the signals or should reflect the meaning that the person giving the signals and doing the behaviour intended. Linked to that is the concern that models should be able to adapt to the individual and cultural differences that influence the interpretation. For example, for a British person, a raised index and middle finger with the palm facing in is a vulgar hand gesture. The same gesture for a North American simply represents the number 2.

4. Conclusion

This paper has tried to indicate the range of variables, modalities, relationships, levels, interpersonal differences, contextual effects and interactions that a comprehensive framework needs to accommodate. The next important step is to develop a model that can inform computational analysis the human process of detecting, interpreting and producing social signals and behaviour. The most sensible way to do this is by applying the model to a specific social phenomenon like politeness. By doing so, a model that is both conceptually and statistically driven can be adapted to explain the communicative process of politeness, and then generalized to other communicative processes.

Acknowledgement Preparation of this paper was supported by FP7 projects SEMAINE and SSPnet

References

[1] S. Koch, Psychology: A study of a Science. New York: McGraw-Hill 1959

[2] P. Ekman. An argument for basic emotions. Cognition and Emotion, 6:169-200, 1992.

[3] P. Ekman. (1980). The face of man: Expressions of universal emotions in a New Guinea Village. New York: Garland STPM Press.

[4] Ekman, P. & Friesen, W. V. (1975). Unmasking the face. A guide to recognizing emotions from facial clues. Englewood Cliffs, New Jersey: Prentice-Hall.

[5] Ekman, P. & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, Calif.: Consulting Psychols Press.

[6] G. Celani, M.W. Battacchi, and L. Arcidiacono, L. The understanding of emotional meaning of facial expressions in people with autism. Journal of Autism and Developmental Disorders, 20:57-66, 1999.

[7] W. Gaebel, and W. Wolwer. Facial expression and emotional face recognition in schizophrenia and depression. European Archives of Psychiatry and Clinical Neuroscience, 242:46-52, 1992.

[8] S.G. Hofmann, M. Suvak, and B.T. Litz. Sex differences in face recognition and influence of facial affect. Pers and Individual Differences, 40:1683-1690, 2006.

[9] S.C. Widen, and J.A. Russell. Children acquire emotion categories gradually. Cognitive Devel, 23:291-312, 2008.
[10] H.H. Clark. Using language. Cambridge: CUP, 1996.
[11] P. Brown, and S.C. Levinson. Politeness: Some universals in language usage. Cambridge: Cambridge University Press, 1987.
[12] K. Grammer, V. Fivola, and M. Fieder. The communication paradox and possible solutions: towards a radical empiricism, in A. Schmitt, K. Atzwanger, K. Grammer, and K. Schäfer, New aspects of human ethology, pages 91–120. Plenum Press, 1997.
[13] K. Carlson, M.W. Dickey, L. Frazier, and C. Clifton, Jr. Information structure expectation in sentence comprehension. The Quarterly Journal of Experimental Psychology, 62:114-139, 2009.
[14] E. Ignatius, and M. Kokkonen. Factors contributing to verbal self-disclosure. Nord Psychol, 59:362-391, 2007.
[15] P. Rockwell. Vocal features of conversational sarcasm: A comparison of methods. Journal of Psycholinguistic Research, 36:361-369, 2007.
[16] P.M. Pexman, T.R. Ferretti, and A.N. Katz. Discourse factors that influence online reading of metaphor and irony. Discourse Processes, 29:201-222, 2000.
[17] A.W. Young, D. Rowland, A.T. Calder, N.,L Etcoff, A. Seth, & Perrett, D.T. Facial Expressions megamix: Tests of dimensional and category accounts of emotion recognition. Cognition 63: 271-313. 1997
[18] L. Reichenbach & J.D. Masters. Children's use of expressive and contextual cues in judgements of emotion. Child Development, 54: 993-1004, 1983.
[19] Scherer, K. R. & Ellgring, H. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? Emotion 7:113–130, 2007.Ambadar, Z., Cohn, J. F., & Reed, L. I. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of Nonverbal Behavior, 33:17-34, 2009.
[20] S. Baron-Cohen, S. Wheelwright, and T. Jolliffe. Is there a "language of the eyes"? Evidence from normal adults, and adults with autism or Asperger Syndrome. Visual Cognition, 4,:311-331, 1997.
[21] S.L. Tasker and L.A. Schmidt. The 'Dual usage problem' in the explanations of 'joint attention' and children's socioemotional development: A reconceptualization. Develop Rev, 28:263-288, 2008.
[22] M. Rowlands. (2006). Body language: Representation in action. Cambridge, US: MIT Press.
[23] J.E. Hansen, and W.J. Schuldt. Physical distance, sex, and intimacy in self-disclosure. Psych Reports, 51:3-6, 1982.
[24] P.J. de Jong. Communicative and remedial effects of social blushing. J. Nonverbal Behav, 23:197-217, 1999.
[25] B.R. Burleson, A.J. Holmstrom, and C.M. Gilstrap. "Guys can't say that to guys": Four experiments assessing the normative motivation account for deficiencies in the emotional support provided by men. Communication Monographs, 72:468-501, 2005.
[26] C. Greenleaf, H. Chambliss, D.J. Rhea, S.B. Martin, and J.R. Morrow Jr. Weight stereotypes and behavioral intentions toward thin and fat peers among white and Hispanic adolescents. J of Adolescent Health, 39:546-552, 2006.
[27] Dawkins, R., & Krebs, J. R. (1978). Animal signals: information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), Behavioral ecology: An evolutionary approach (2nd Ed.), pp. 282-309. Blackwell Scientific.
[28] Dunbar, R. I. M. (1996). Grooming, Gossip and the Evolution of Language. London: Faber and Faber.
[29] Knight, C., (2000). The evolution of cooperative communication. In Knight, C., M. Studdert-Kennedy & J. R. Hurford (eds), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form. Cambridge: CUP, pp. 19-26.
[30] R Cowie (in press) Describing the forms of emotional colouring that pervade everyday life In P Goldie (ed) The Oxford Handbook of the Philosophy of Emotion: Oxford: Oxford University Press
[31] W.L. Cook, and D.A. Kenny. The actor-partner interdependence model: A model of directional effects in developmental studies. International Journal of Behavioral Development, 29:101-109, 2005.
[32] M. Battaglia, A. Ogliari, A. Zanoni, F. Villa, A. Citterio, F. Binaghi, et al. Children's discrimination of expressions of emotions: Relationship with indices of social anxiety and shyness. J of American Acad of Child and Adolescent Psychiatry, 43:358-365, 2004.
[33] L.A. Melchoir, and J.M. Cheek. Shyness and anxious self-preoccupation during a social interaction. Journal of Social Behavior and Personality, 5: 117-130, 1990.
[34] D.J. Bem, and A. Allen. On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psych Rev, 81:506-520, 1974.
[35] P.M. Brunet, and L.A. Schmidt. Is shyness context specific? Relation between shyness and online self-disclosure with and without a live webcam in young adults. J of Research in Personality, 41, 938-945, 2007.
[36] M.E. Ainsfield, B.M. DePaulo, and K.L. Bell. Familiarity effects in nonverbal understanding: Recognizing our own facial expressions and our friends'. Journal of Nonverbal Behavior, 19:135-149, 1995.

The Action Synergies: Building Blocks for Understanding Human Behavior

Yi Li
Electrical and Computer Engieering
University of Maryland, College Park
liyi@umiacs.umd.edu

Yiannis Aloimonos
Computer Vision Lab
University of Maryland, College Park
yiannis@cfar.umd.edu

Abstract

Social signal processing is an emerging field that gains more and more attention. As a key element in the field, visual perception of human motion is important for understanding human behavior in social intelligence. Motivated by the hypothesis of muscle synergies, we proposed action synergies for automatically partitioning human motion into individual action segments in videos. Assuming the size of the human subject is reasonable and the background changes smoothly, the video sequence is represented by six latent variables, which we obtain using Gaussian Process Dynamical Models (GPDM). For each variable, the third order derivative and its local maxima are computed. Then by finding the consistent local maxima in all variables, the video is partitioned into action segments. We demonstrate the usefulness of the algorithm for periodic motion patterns as well as non-periodic ones, using videos of various qualities. Results show that the proposed algorithm partitions videos into meaningful action segments.

1. Introduction

Social signal processing has a promising potential for exploring the basic problems in social intelligence. By trying to discover the structures of the signals widely available from visual space to motor space, it is possible to unveil how humans react and interact. Language is naturally involved as the description to the signals in these two spaces. Altogether, these three spaces create a powerful triangle for understanding the social intelligence (Fig. 1).

By understanding the problems in each space individually and the mapping from one space to the other, we clearly define and address the fundamentals in social signal processing. For example, mirror neuron suggested that the visual-motor primitives are important in understanding the cognitive issues in the field. If one can interpret the signals in visual space (video) using the control signal in motor space, a much better humanoid robot will have the power of mining reality in the society. Take the mapping between the language space and the motor space as another example. This mapping gives us the potential to investigate the underlying principles of the body language.

Primitives serve as the bridge between the low level signal and the high level description in social intelligence. To discover the primitives in any space, our point of view is that the social signal should be decomposed into building blocks. Based on these building blocks, one can further study the primitives in visual/motor space, and parse human actions in a symbolic way [12]. In this paper, we focus on automatically partitioning human motion in video space into small yet meaningful action segments. This makes it possible to further reveal the primitives in visual space and map them to the motion control signal.

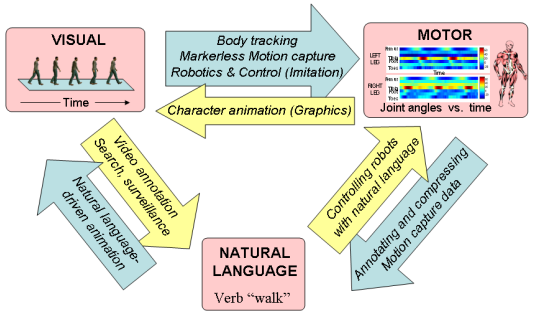


Figure 1. Three spaces in social signal processing. Clockwise from upper left, the signals in the spaces are videos, motion capture (MoCap) data, and discrete symbols, respectively.

A natural "by product" of the segments are the first and the last pose of the action. We name them "representative poses" because in many scenarios they are also good for action recognition using pose estimation techniques. Note, it is not clear how human defines representative poses in actions, but it is very clear that between two representative actions there should be one complete action (e.g., moving from one bar to another in Fig. 2). Such a representation can serve as basis for further action analysis using both the video segments and the representative poses. In addition, this representation makes it possible to study non-periodic motion patterns as well as periodic ones.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

5 18 32 41

Figure 2. Representative poses of the video "girl on the monkey bar". The video is available at Youtube. The frame numbers are displayed below. Note that the length of each action instance varies even when the action is periodic.

How can we find these segments? d'Avella *et al.* [8] discovered the concept of muscle synergies which suggests that groups of muscles are coherently activated. This indicates that a motion sequence can be partitioned to segments at the time when body parts have a consistent change in acceleration.

The idea extends to vision. Certainly we cannot measure the muscle signal from video, but empirically we are able to transfer the above idea to the visual domain and use them to segment actions by performing dimensionality reduction. Most probably because of the underlying muscle synergies, the extrema in the change of acceleration of the signals in the reduced space are happening about at the same time. This suggests that a motion sequence can be partitioned to segments at the time when body parts have a consistent change in acceleration. In addition, the poses at these breaking points can be considered as the representative poses for the segments. We call this the "action synergies".

The paper is organized as follows. Sec. 2 discusses the related work. Sec. 3 presents the algorithm for partitioning human motion video. Sec. 4 shows the experiments and comparisons, and Sec. 5 concludes the paper.

2. Related work

Actions exist both in motor space (actions we do) and in visual space (actions we see). We describe here prior art in both spaces.

Finding primitives of human actions in motor space is the basic topic in different fields, where stick-figure models (skeleton) are used. The temporal segmentation, manual or automatic, of joint motion is typically the first step. Jenkins and Mataric [13] used KCS, a heuristic algorithm, to partition the motion. Lu and Ferrier [16] assumed that joint motion is an autoregressive process and partitioned the data based on different process parameters. A comparison of some partitioning algorithms in motor space can be found in [5], but only periodic motion patterns or a small number of actions [18] are considered. Alternatively, Guerra-Filho and Aloimonos [12] used the sequences of consecutive joint angles which have the same sign of velocity and acceleration as the primitives in motor space. Zhou *et al.* [30] used k-mean clustering and the Component Analysis technique to partition the joint angle series into different action groups.

There is a large body of work on recognizing actions in visual space ([4,11,14,15,22,23]). 2D features from image sequences, such as optical flow, silhouette, and similarities between frames, are computed and mapped directly to semantic concepts such as walking [2], running [9] and dancing [20]. Alternatively, 2D features from image sequences are mapped to the 2D joint space [21]. The body joints can be further analyzed in the 2D joint space [17]. Sigal and Black [24] built a human motion dataset for evaluation, tracking, and pose estimation.

Partitioning actions in video has not been addressed in the study of action recognition. Video temporal segmentation has been mainly used for shot boundary detection, key frame extraction [1], video content analysis [27], and video synopsis [19]. Such segmentations are helpful for retrieving useful information from a large collection of videos, but these techniques do not aim at partitioning human motion into action segments. Action synopsis [3] is related to our work but the joint locations need to be labeled using a semi-automatic software called Icarus prior to the analysis.

Dimension reduction techniques have been used to analyze human motion both in motor space [7, 25] and in visual space [10,26]. Silhouettes of periodic patterns, such as walking and running, are extracted and embedded in a latent space. Studies show closed curves for periodic patterns in latent spaces, but the partitioning is not addressed.

In related work, neuroscientists also study human motion in a physically meaningful way. Velocity, acceleration, and jolt are the measurements for input motion streams [29].

3. Partitioning human motion in visual space

We present the action partitioning algorithm in this section. A human motion video is represented by the GPDM variables in a low dimensional latent space. The time series of each latent variable implicitly encodes the process that generates the human motion. The third order derivative of the time series, which we called "action synergies" in the paper, is computed for each latent variable. We observe that the local maxima of the action synergies approximately correspond to the representative poses. By finding the consistent local maxima in the action synergies of the different variables, we partition the video.

3.1. Representing video in the GPDM latent space

The Gaussian Process Dynamical Models (GPDM) [25] assumes that the data in high dimension can be compressed to a latent space using Gaussian priors for both stochastic dynamical process and mapping. In [25], the GPDM maps the joint angles in human MoCap data to a 3D latent space.

The relation between the latent variables $\mathbf{X}$ and the original data $\mathbf{Y}$ is as follows:

$$p(\mathbf{X}, \mathbf{Y}, \overline{\alpha}, \overline{\beta}, W) = p(\mathbf{Y}|\mathbf{X}, \overline{\beta}, W)p(\mathbf{X}|\overline{\alpha})p(\overline{\alpha})p(\overline{\beta})p(W) \quad (1)$$

where $\overline{\alpha}$, $\overline{\beta}$, and W are the parameters. Given $\mathbf{Y}$, one estimates the embedding $\mathbf{X}$ in the reduced space.

GPDM models the uncertainty and sparsity in $\mathbf{Y}$ using Gaussian priors, learns the effective representation of the nonlinear dynamics in high dimensional spaces, and uses a small number of variables $\mathbf{X}$ as the representation of the original signal in the reduced space. $\mathbf{X}$ can be further interpreted as the underlying variables that govern the signal in the high dimensional space. Refer to [25] for details.

In the proposed algorithm, we apply the GPDM directly on the images. A high dimensional vector of image intensities is formed by concatenating the pixel value from left to right and then from top to bottom in each image. The GPDM embeds the sequence of intensity vectors in a low dimensional latent space.

In our experiment, we expect to use a minimal number of variables that can capture the human motion dynamics without losing much information. We tested various number of latent dimension, and experimentally selected 6D because it generally sufficient in our experiments.

As an example, Fig. 3 visualizes the trajectory of the latent variables for the video shown in Fig. 2 in the GPDM dimensions 1-3 and 4-6, respectively. The latent variables in the reduced space might not have concrete physical meanings but they are indirectly linked to the underlying muscle signals that generate the videos.

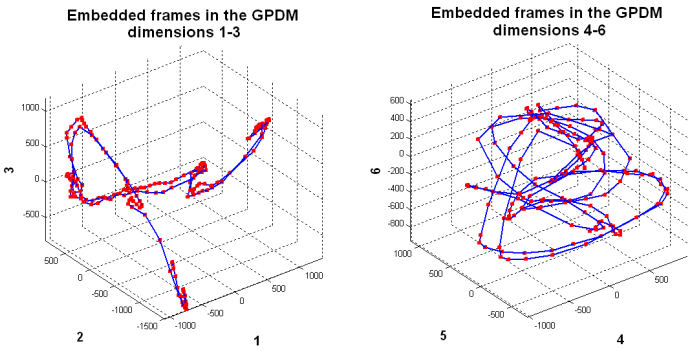


Figure 3. The embedded frames (red) and the trajectory of the latent variables (blue) in the GPDM dimensions 1-3 (left) and 4-6 (right), respectively. The input is the video shown in Fig. 2 (186 frames in total).

3.2. The action synergies

We compute the action synergies for each latent variable. The latent variables, $\mathbf{X}(t)$, represent the embedded frame at time t in the 6D GPDM space

$$\mathbf{X}(t) = [x_1(t), x_2(t), ..., x_6(t)]^T \quad (2)$$

Each latent variable, $x_i(t)$ where $i = [1...6]$, is a 1D time series (Fig. 4a). The jolt of $x_i(t)$ is computed as

$$J_i(t) = \frac{d^3(x_i(t))}{dt^3} \quad (3)$$

To minimize the computational error, we use the 7 point algorithm followed by low pass filtering to smooth the data.

To measure the jolt in a better way, we compute the jolt envelope (Fig. 4b) as

$$\text{Env}_i(t) = ||\text{Hilbert}(J_i(t))|| \quad (4)$$

for $i = [1...6]$, where Hilbert$(\cdot)$ is the Hilbert transform and $||\cdot||$ is the L_2 norm. This is a standard approach to computing signal envelope [6]. Then we use a Butterworth filter as a low pass filter to process $\text{Env}_i(t)$, and compute the envelope peaks (local maxima) of the filtered $\text{Env}_i(t)$ for $i = [1...6]$ (Fig. 4d).

We observe that the envelope maxima of different latent variables cluster. Each row in Fig. 5 (shown from a perspective viewpoint) indicates the locations of the envelope peaks in Fig. 4b for each latent variable. The frames corresponding to the centers of the clusters across different dimensions are shown on top. One can see that the locations of envelope peaks approximately correspond to similar poses for repetitive actions.

This observation indicates the "breaking points" in the motion stream can be found by selecting the consistent envelope peaks of different latent variables (e.g., the blue bars in Fig 5). This can be solved as an optimization problem.

For comparison, we also use the Principal Component Analysis to process the video. Results are shown in Fig. 4c and 4d, respectively. They show that the clusters of the envelope peaks of the GPDM latent variables (Fig. 4b) are more consistent than those of the PCA variables (Fig. 4d).

In our method, only local maxima of the change in acceleration need to be finally computed. Therefore, the above observation is visible for a reasonable range of smoothing parameters' values. Consequently, the optimization procedure in Sec. 3.3 is loosely coupled with these parameters.

3.3. Partitioning videos by finding consistent envelope peaks in different latent variables

A consistent envelope peaks is the collection of the envelope peaks in different latent variables with the minimal sum of the pairwise distances. The video is partitioned into two halves at the location of the center of the envelope peaks. This procedure is iteratively applied on each half, and in this way the video is partitioned into action segments.

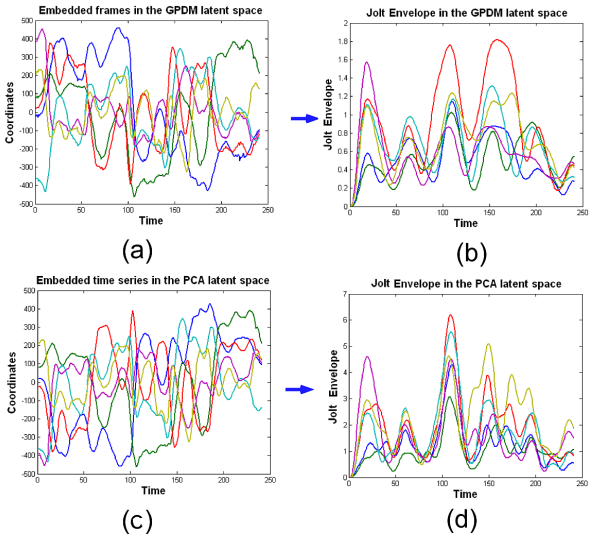


Figure 4. The embedded frames (a and c) and the action synergies (b and d) in the 6D GPDM latent space and in the PCA space, respectively. Different colors represent the time series for different latent variables. The input video is the "Lena on the monkey bar" (Fig. 5a).

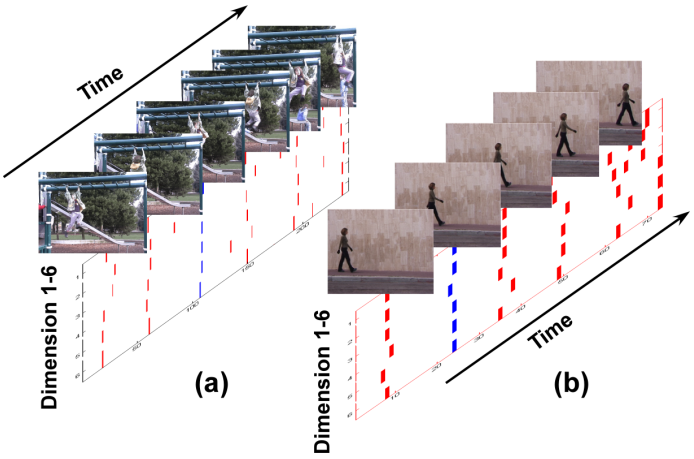


Figure 5. Illustration of the algorithm. Each row of the graphics in (a) shows the locations of the envelope peaks in Fig. 4b for each latent variable. The frames, which correspond to the centers of the clusters, are shown on top. By selecting the envelope peaks in different variables with minimal sum of the pairwise distances (blue bars), the video is partitioned into two halves. (a) "Lena on the monkey bar" sequence; (b) a "walking" sequence from the Weizmann dataset.

We formulate this procedure as an optimization problem. The goal is to choose one of the envelope peaks from each latent variable such that the sum of the pairwise distances is minimal. Therefore, we use an indicator function δ_m^i for the location of the i^{th} envelope peak of the m^{th} latent variable I_m^i such that

$$\delta_m^i = \begin{cases} 1, & \text{if } I_m^i \text{ is selected} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

for all possible i, and

$$\sum_i \delta_m^i = 1 \tag{6}$$

for $m = [1, ..., 6]$.

The distance between the locations of two envelope peaks I_m^i and I_n^j $(m \neq n)$ is defined as

$$d(I_m^i, I_n^j) = \exp(\frac{1}{\sigma}|I_m^i - I_n^j|) - 1 \tag{7}$$

where σ is the parameter for the distance measurement.

Then, the cost of the consistency between a set of envelope peaks in different latent variables is the sum of the pairwise distances,

$$c = \sum_m \sum_{n,n\neq m} \sum_i \sum_j \delta_m^i \delta_n^j d(I_m^i, I_n^j). \tag{8}$$

Now, the optimization problem is to find δ_m^i which minimize c,

$$\delta^* = \arg\min_\delta c \tag{9}$$

subject to Eq. 6.

Each iteration gives two segments, then the partitioning algorithm is iteratively applied on each segment until the value of C is smaller than a threshold κ.

4. Experimental demonstration of the usefulness of the segmentation algorithm

We use six videos from four different categories in the experiments: 1) three videos from public datasets (one "walking" and one "jumping" from the Weizmann dataset, one "jogging" from the KTH dataset, respectively). The backgrounds of these videos are (almost) static; 2) two public accessible videos from Youtube. One is an indoor sequence "man in exercise"[1], and the other is an outdoor sequence "girl on the monkey bar" (Fig. 2). These videos are taken with moving cameras; 3) one black and white historical video clip from a slowly moving camera[2]. A part of the video is used for this experiment;

In all the experiments, consecutive frames containing representative poses are shown. We resize all the images to have the same height (50 pixels) while preserving the height/width ratio. The parameters for filtering and the minimal threshold for consistency depend on the frame rate.

4.1. Results

First we present the results for the sequences from public datasets. Three videos, "walking", "jogging", and "jump-

[1] www.youtube.com/watch?v=57I4-QXCtRA
[2] www.youtube.com/watch?v=iv6p9XbIhtA

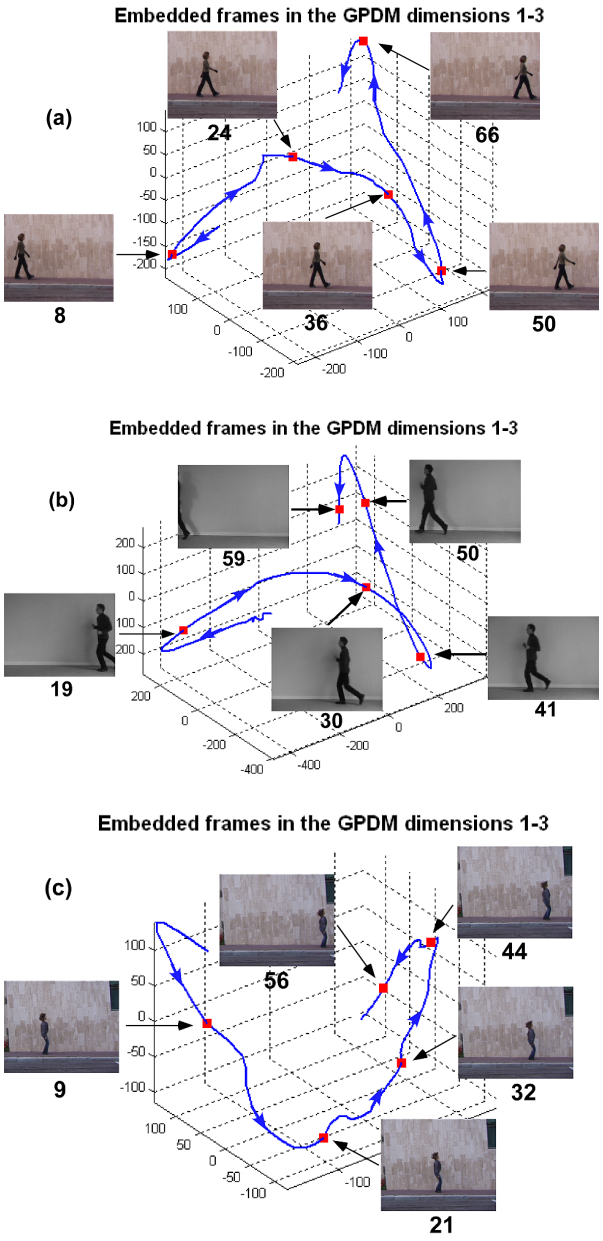


Figure 6. Results for the action videos from public datasets. (a) "walking"; (b) "jogging"; (c) "jumping". For each action sequence, the trajectory of the latent variables and the representative poses are shown in the GPDM dimension 1-3. Frame numbers are displayed below the representative poses. The blue arrow denotes the time direction.

ing" are used for demonstration. Fig. 6 shows the trajectory of the embedded frames in the GPDM dimension 1-3, as well as the estimated representative poses. The blue arrow denotes the time direction. In other approaches, silhouettes are extracted first. We do not require the image segmentation. Nevertheless, the representative poses in test videos are correctly identified. Each segment between two representative poses was found exactly one action, i.e., "one step forward" in Fig. 6a and 6b, and "one jump" in Fig. 6c.

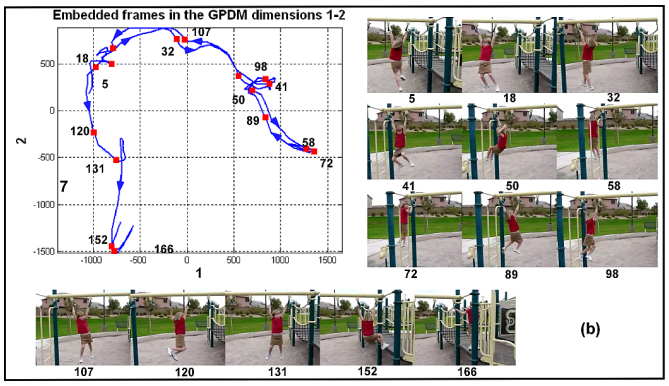


Figure 7. Results for the action videos from Youtube (with permission from the video owners). (a) "man in exercise"; (b) "girl on the monkey bar". For each action sequence, the trajectory of the latent variables is shown in the GPDM dimension 1-2. Frame numbers are displayed under the representative poses as well as in the GPDM space. The blue arrow denotes the time direction.

The static background in Fig. 6 makes the GPDM easy to learn the parameters. In the second experiment, we used the video "man in exercise", which was captured with a moving camera in a gymnastic room (Fig. 7a). Representative poses are correctly identified even though the viewpoint changes and the background changes smoothly as well.

The motions in the previous four videos are periodic and the time differences between two representative poses are similar. Fig. 7b ("girl on the monkey bar"), taken with a moving camera, shows a more challenging situation where the time the girl moves from one bar to another varies, and the variation between action segments is large. The partitioning result shows that we capture the poses when the girl moves from left to right and vice versa. Frame 151 and 166 in Fig. 7b show that we captured the "landing" on the platform correctly.

An even more challenging situation is the historical black and white video of a martial artist captured with a slightly moving camera (Fig. 8). The human motion is non-periodic, and the quality of this video is very low

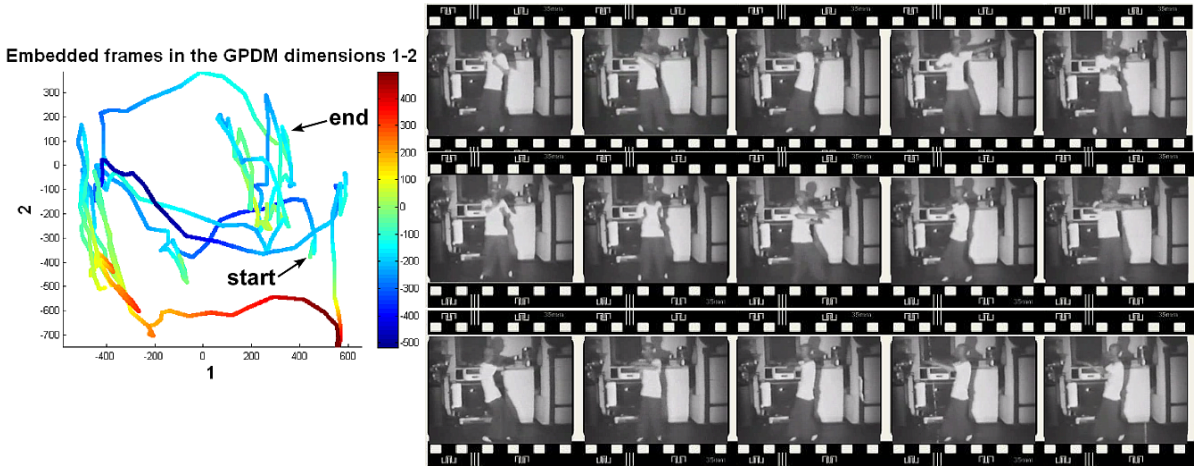


Figure 8. Result for a low quality black and white video from Youtube. The master practices "Wing Chun", a martial arts style. The trajectory of the latent variables in the GPDM dimensions 1-2 are shown in the left. The color represents the value of the latent variable in the GPDM dimension 3. The first frame ("start") and the last frame ("end") are indicated in the latent space.

(e.g., the shadow could impair the performance of the foreground/background segmentation). Nevertheless, Fig. 8 shows that the representative poses are clearly identified as the final poses of separate movements.

The current algorithm shows very promising results on these challenging situations. Results for videos of various qualities demonstrate that each segment contains one complete action and the representative poses are useful for understanding the actions.

4.2. Preliminary evaluation result

We present our preliminary evaluation result in this section. We manually labeled all the videos shown in Fig. 6, Fig. 7, and Fig. 8. In each video, we labeled the frame where the human subject changed from one action to another as the "breaking points". Then we used the Hungarian algorithm to evaluate the performance of different algorithms compared to the ground truth.

First, we compute the optimal one-to-one-mapping between the estimated intervals and the ground truth intervals. Then the partitioning accuracy is defined as the sum of the overlap between the corresponding pairs divided by the total length of the intervals. This "assignment" problem can be solved by the Hungarian algorithm (see [28] for details). Table 1 shows that the GPDM has a better assignment score than all other algorithms when compared to the ground truth.

Table 1. Accuracy of the partitioning using different dimension reduction algorithms. We manually labeled the videos in Fig. 6, Fig. 7, and Fig. 8 as the ground truth. The accuracy rate is computed using the Hungarian algorithm.

Algo	GPDM	Isomap	Kernel PCA	Laplacian	LLE
Rate	0.94	0.90	0.85	0.86	0.84

5. Conclusion

We proposed an algorithm for partitioning human motion video into action segments, one action per segment. Our basic idea is the action synergies and the consistent occurance of local maxima in the third order derivatives along the dimensions of a latent space. Experiments demonstrate that our algorithm is useful for partitioning the video into meaningful action segments. Our future work includes finding the primitives from common actions in visual space using the action segments.

References

[1] *http://www-nlpir.nist.gov/projects/trecvid*.

[2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(1):44–58, 2006.

[3] J. Assa, Y. Caspi, and D. Cohen-Or. Action synopsis: pose selection and illustration. *ACM Trans. Graph.*, 24(3):667–676, 2005.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV '05*, pages 1395–1402, 2005.

[5] D. Bouchard. *Automated Motion Capture Segmentation using Laban Movement Analysis*. PhD thesis, University of Pennsylvania, 2008.

[6] R. Bracewell. *The Fourier Transform & Its Applications*. McGraw-Hill Science, June 1999.

[7] R. Chalodhorn, D. Grimes, G. Maganis, R. Rao, and M. Asada. Learning humanoid motion dynamics through sensory-motor mapping in reduced dimensional spaces. In *ICRA'06*.

[8] A. d'Avella, P. Saltiel, and E. Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nature Neuroscience*, 6:300–308, 2003.

[9] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. pages 726–733 vol.2, 2003.

[10] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR (2)*, pages 681–688, 2004.

[11] S. Gong, J. N. S. Kwong, and J. Sherrah. On the semantics of visual behaviour, structured events and trajectories of human action. *Image Vision Comput.*, 20(12):873–888, 2002.

[12] G. Guerra-Filho and Y. Aloimonos. A language for human action. *IEEE Computer*, 40(5):42–51, 2007.

[13] O. Jenkins and M. Mataric. Deriving action and behavior primitives from human motion data. In *IROS'02*, pages 2551–2556.

[14] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. *CVPR'07*.

[15] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. *CVPR'07*.

[16] C. Lu and N. Ferrier. Repetitive motion analysis: Segmentation and event classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(2):258–263, 2004.

[17] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *International Journal of Computer Vision*, 66(1):83–101, 2006.

[18] V. Pavlovic, J. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *NIPS*, pages 981–987, 2000.

[19] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1971–1984, 2008.

[20] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. A. Viola. Learning silhouette features for control of human motion. *ACM Trans. Graph.*, 24(4):1303–1331, 2005.

[21] R. Rosales and S. Sclaroff. Learning body pose via specialized maps. In *NIPS*, pages 1263–1270, 2001.

[22] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *CVPR'08*.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR'04*, volume 3, pages 32–36 Vol.3, 2004.

[24] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity Technical Report*, 2006.

[25] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(2):283–298, 2008.

[26] X. Wang, L. Wang, and A. Wirth. Pattern discovery in motion time series via structure-based spectral clustering. In *CVPR*, 2008.

[27] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):893–908, 2008.

[28] L. Yi, Y. Zheng, D. Doermann, and S. Jaeger. Script-Independent Text Line Segmentation in Freestyle Handwritten Documents. *IEEE Trans. Pattern Anal. Machine Intell.*, 2008.

[29] V. Zatsiorsky. *Kinematics of Human Motion*. Human Kinetics Publishers, September 1997.

[30] F. Zhou, F. Frade, and J. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *IEEE Conference on Automatic Face and Gestures Recognition*, September 2008.

Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations?

Katrin Solveig Lohan Anna-Lisa Vollmer Jannik Fritsch Katharina Rohl ng
Britta Wrede
CoR-Lab, Applied Informatics Group
Bielefeld University, Bielefeld, Germany
http://www.cor-lab.de
klohan@techfak.uni-bielefeld.de

Abstract

In developmental research, tutoring behavior has been identified as scaffolding infants' learning processes. Infants seem sensitive to tutoring situations and they detect these by ostensive cues [4]. Some social signals such as eye-gaze, child-directed speech (Motherese), child-directed motion (Motionese), and contingency have been shown to serve as ostensive cues. The concept of contingency describes exchanges in which two agents interact with each other reciprocally. Csibra and Gergely argued that contingency is a characteristic ostensive stimulus of a tutoring situation [4]. In order for a robot to be treated similar to an infant, it has to both, be sensitive to the ostensive stimuli on the one hand and induce tutoring behavior by its feedback about its capabilities on the other hand.
In this paper, we raise the question whether a robot can be treated similar to an infant in an interaction. We present results concerning the acceptance of a robotic agent in a social learning scenario, which we obtained via comparison to interactions with 8-11 months old infants and adults in equal conditions. We applied measurements for motion modifications (Motionese) and eye-gaze behavior. Our results reveal significant differences between Adult-Child Interaction (ACI), Adult-Adult Interaction (AAI) and Adult-Robot Interaction (ARI) suggesting that in ARI, robot-directed tutoring behavior is even more accentuated in terms of Motionese, but contingent responsivity is impaired. Our results confirm previous findings [14] concerning the differences between ACI, AAI, and ARI and constitute an important empirical basis for making use of ostensive stimuli as social signals for tutoring behavior in social robotics.

1. Introduction

In social learning, infants bene t from the behavior of their tutors. The modi ed behavior seems to help infants to lter the information that is crucial for learning. Csibra and Gergely [4] highlight the importance of this pedagogic behavior that is crucial for the understanding of some actions: pedagogy essentially created a new way of information transfer among individuals through the use of ostensive communication . In their work, they give the example of peeling a hard fruit or carve away pieces of wood with a tool. The movement and the tool in both actions are the same, but the goal and reason for the action are very different. Where it is easy to infer the goal of the action when peeling a fruit, i.e. getting to the edible parts, it is not obvious what is intended in the case of the wood carving. Therefore, tutoring is crucial in order for a learner to understand the goal correctly. Csibra and Gergely [4] argue that economical reasons account for tutoring, because otherwise learning would not be feasible. Tutoring situations thus are created by the tutor via ostensive stimuli, which are originally evolved to assist pedagogy . The effect of pedagogy seems to rely on the bidirectionality. Csibra and Gergely (2005) explain the contribution achieved by the learner, who has to send signals during the course of tutoring telling the tutor when s/he is attentive and receptive and possibly showing understanding. Furthermore, infants seem sensitive to tutoring situations and ostensive cues help them to detect these [13]. The term ostensive cues refers to social signals such as eye-gaze, child-directed speech (Motherese) [5], child-directed motion (Motionese) [2,6,7], and contingency [4]. While the phenomenon of multimodal child-directed speech (Motherese) or action (Motionese) is widely known, the concept of contingency is less popular. It describes exchanges in which two agents interact with each other reciprocally. Csibra and Gergely ([4], p.8) argue that contingent responsivity is a characteristic ostensive stimulus of a tutoring situation: If a source repeatedly appears to remain silent during your actions but starts to emit signals as soon as you have stopped your actions, it gives

978-1-4244-4799-2/09/$25.00

you the strong impression that the source is communicating with you . The idea of creating a robot that actively lters the information from the environment and manages to attend to certain sources of information while ignoring others has to be supported by the robot's sensitivity to the ostensive stimuli on the one hand and induce tutoring behavior by its feedback about its capabilities on the other hand. A robot which has the appearance of an infant should hence be able to pro t from these behavior modi cations as well. Recently, Vollmer et al. found that adults modify their behavior when interacting with children (ACI) and robots (ARI) as opposed to adult-directed interaction (AAI) [14]. Modi cations were found with respect to Motionese measurements, indicating that in ACI and ARI movements were slower, less round and had a slower pace than in AAI indicating that subjects behave similar towards robots and infants. However, number and length of eye-gaze bouts differed signi cantly between ACI and ARI with less eye-gaze bouts and less long eye-gaze bouts directed towards the interaction partner in ARI. This indicates that contingency was impaired in the ARI condition. In this paper, we report on results from a task with a similar structure based on a more ne grained analysis of the eye-gaze behavior in order to

- show how far the ndings by Vollmer et al. hold for a different task
- analyze the structure of eye-gaze behavior over time and
- discuss these results with respect to the question in how far the observed modi cations of behavior can be interpreted as ostensive signals in human-robot interaction.

2. Experiment

Two experiments were carried out to obtain data from parent-infant and adult-robot interactions [14]. The data on adult-child interaction is based on the same setting as in [12] and [10]. The data on human-robot interaction was obtained in a second experiment as described in [14]. From the overall set of items that were presented we selected the Minihausen task. This task is similar to the stacking-cups task as it is a rather goal-directed action with three sub-goals to be reached. Results from analyses of motionese and contingency features in parent-infant and adult-robot interaction have shown that while motionese features of infant-directed and robot-directed interactions are similar, they diverge for contingency measures, indicating that contingency is impaired in human- robot-interaction, [14]. In this paper we ask the question in how far these results are decisive for the statement that motionese as well as contingency features serve the function of ostensive signals.

2.1. Motionese Experiment (ME)

2.1.1 Subjects

The Motionese Corpus consists of infant- and adult-directed interactions. We selected the younger group comprising 12 families of 8 to 11 months old children. Both parents were asked to demonstrate functions of 10 different objects to their children as well as to their partners or another adult. In the following, we focus on the analysis of the Minihausen task, because it offers good comparability in motion performance. We further selected a subgroup of 8 parents (4 fathers and 4 mothers) for the ACI and a subgroup of 12 parents (7 fathers and 5 mothers) for the AAI, because of the quality of the video, sound and due to the way in which the action was performed. More speci cally, the order in which the blocks of the considered Minihausen task are put onto the wooden base poles can vary: We selected only those parents, who started the task by putting the rst block -the one closest to the body- onto the respective pole which means putting the blue block onto the rightmost pole. (see Fig. 3 a1).

2.1.2 Setting

Parents were instructed to demonstrate a Minihausen task to an interaction partner. The interaction partner was rst their infant and then an adult. Fig. 1 illustrates the top-view of the experimental setup. The Minihausen task was to sequentially pick up the blue (a1), the yellow (a2), and the green (a3) block and put them onto the wooden base with three poles on the white tray.

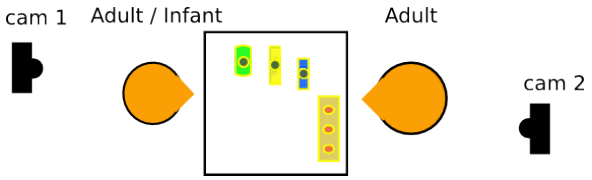


Figure 1. Motionese Setting, there are two cameras which are recording the scene. The interaction partners are seated across from each other and the object is laid on the table in front of the tutor.

2.2. Robot-Directed Interaction Experiment (RDIE)

2.2.1 Subjects

31 adults (14 females and 17 male) participated in this experiment 7 out of which were parents as well. Out of this group, we selected 12 participants (8 female and 4 male), who performed the task in a comparable manner.

2.2.2 Setting

The participants were instructed to demonstrate several objects to an interaction partner, while explaining him/her how to do it (Fig. 2). Again we chose the Minihausen task for analysis. The interaction partner was an infant-like looking virtual robot with a saliency-based visual attention system [10]. The robot-eyes will follow the most salient point in the scene, which is computed by color, movement, and other features (see [10]).

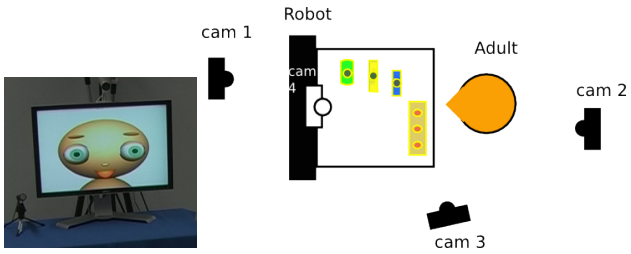


Figure 2. The robot simulation presented on the screen can be seen on the left picture. The rigth picture shows the Robot-directed Interaction Setting, there are four cameras which are recording the scene. The subject is seated across from the robot and the object is laid on the table in front of the tutor.

3. Data Analysis

The goal of this paper was to analyze those cues, that we hypothesize to serve as social signals in tutoring behavior. These can be grouped into two groups, one that measures Motionese and another one that that may be used to measure Contingency. We coded the videos semi-automatically to obtain data for the 2D hand trajectories and the eye gaze directions.

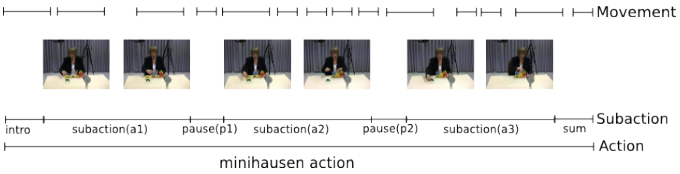


Figure 3. The action was devided into movment and pause parts and into subactions. This graphic shows an example for the structure of an 'Action', 'Subaction'(intro = Introduction and sum = summary), and 'Movement'.

3.1. Annotations

For all annotations, we used the video captured by camera (cam) 1, see Fig. 1 and 2. It shows the front view on the demonstrator and is therefore best suited for action, movement, and gaze annotations, which are discussed in detail below.

3.1.1 Motionese

Action Segmentation: For analyzing the data, the action of the Minihausen task and additionally, the sub-actions (a1-a3) of grasping one block until releasing it onto the end position (Fig. 3) were marked in the video. We de ned

1. action as the whole process of transporting all objects to their goal positions.
2. subaction as the process of transporting one object to its goal position.
3. movement as phases where the velocity of the hand is above a certain threshold. All other phases are de ned as pauses.

Hand Trajectories: The videos of the two experiments were analyzed via a semiautomatic hand tracker system (Fig. 4). The system is written as a plug-in for a graphical plug in shell, iceWing [8], and makes it possible to track both hands with an Optical Flow based algorithm, Lucas & Kanade [9]. It allows manual adjustment in case of tracking deviation. We used this tracking system instead of a previously used 3D body model system, [12], since 3D results in [12] were not signi cant, we focused on 2D analyses which provide to show more stable results. Additionally, the new system is easily accessible for non-expert users.

Figure 4. Example frame for hand tracker system annotation. The red and violet circles depict the tracking regions. The points in the middle of the circles are the resulting 2D points for the hand trajectory.

3.1.2 Contingency

Eye Gaze: In annotating the eye gaze directions with the program Interact [1], we distinguished between looking at the interaction partner, looking at the object and looking anywhere else.

3.2. Measures

For quantifying Motionese and Contingency, we computed ve variables related to the 2D hand trajectories derived from the videos and the eye gaze bout annotations produced with Interact.

3.2.1 Motionese

We measured Motionese in terms of velocity and range as de ned in [14].

Velocity was computed using the derivative of the 2-dimensional hand coordinates of the hand which performed

the action per frame as the average velocity for subactions a1, a2, and a3 each.

Range was de ned for each subaction separately as the covered motion path divided by the distance between motion, i.e. subaction, on- and offset.

3.2.2 Contingency

The Contingency of the interactions was quanti ed in terms of variables related to eye gaze, as de ned in [3] for measuring interactiveness.

The *total length of eye-gaze bouts to interaction partner* de ned as the percentage of time of the action spent gazing at the interaction partner was computed. Brand et al. found that the total length of eye-gaze bouts to the interaction partner in their study was signi cantly greater in ACI than in AAI [3]. Also the *total length of eye-gaze bouts to object* and the *total length of eye-gaze bouts elsewhere* were calculated as the percentage of time of the action spent gazing at the object and somewhere else, as for example at the table or the experimentor.

4. Results

A non-parametric test (Mann-Whitney U test) was run for all pairs of samples, ACI vs. AAI, ACI vs. ARI, and AAI vs. ARI. Table 1 depicts the results of the study.

4.1. Motionese

For the Motionese measures, our results revealed the following:

For the *velocity* measure, which is computed for each subaction and takes into account the hand movement during the transportation of the respective block, the results showed signi cant differences for all three subactions for all pairs of conditions. These results clearly show that in AAI hand movements are faster than in ACI and ARI and additionally that hand movement is slowest in the ARI condition. Also note that for all conditions the mean values increase for the consecutive subactions: velocity in subaction a1 $<$ velocity in a2 $<$ velocity in a3. In ARI, the rate in which the mean values increase is lowest and in AAI the rate is highest. The latter is specially noticable for the last subaction a3.

The *range* measure suggests that ARI exhibits the greatest range for each subaction and therefore movement is most exaggerated. Also, range is greater in ACI than in AAI. For ACI vs. AAI results revealed no signi cance, but a trend for subactions a2 and a3. For ACI vs. ARI solely results for subaction a3 showed signi cance, for a1 and a2 they show a trend. For AAI vs. ARI subactions a2 and a3 revealed signi cance, whereas a1 again shows a trend. Again we can state that in ARI the rst subaction a1 has the highest range value of all subactions over all conditions. Looking at this measure over time, range decreases rapidly to about one half for subaction a2 and some more for the last subaction a3. For the other conditions however the rate of change, i.e. the decrease, is not as drastical.

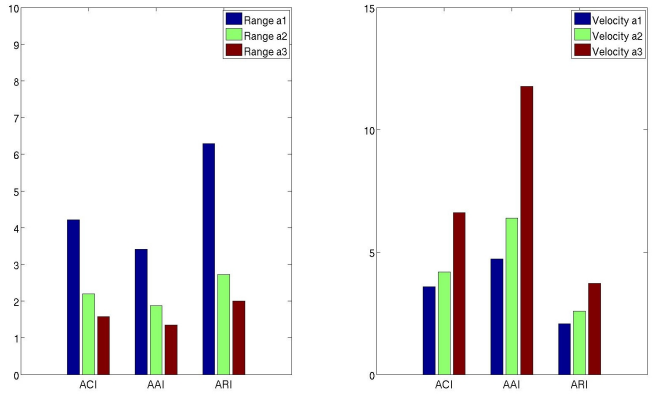


Figure 5. This graph shows the range of hand movement in the three different subactions on the left. On the right, the mean velocity of hand movement in the three different subactions can be seen for the Minihausen -task (y-axis) in every condition (x-axis).

4.2. Contingency

Most interestingly, the results for eye gaze show a completely different picture. For *total length of eye-gaze bouts to interaction partner* they show that in ACI signi cantly more time was spent gazing at the interaction partner than in AAI and ARI. Differences between AAI and ARI are not signi cant. Looking at this measure over time, it is interesting to notice that in all three conditions the most time of gazing at the interaction partner was spent in the summary part of the action, sum.

For the measure *total length of eye-gaze bouts to object*, values are signi cantly lower in ACI than in AAI and ARI, where differences between AAI and ARI exhibit that values are signi cantly lower in ARI.

The *total length of eye-gaze bouts elsewhere*, which measures the percentage of time gazed neither to interaction partner nor object, reveals that most time gazing somewhere else is spent in the ARI condition, followed by ACI. The differences between ACI and AAI could be a result of the design of the study, because the AAI follows the ACI, so that instructions and experimentor are not anymore needed to turn to for help in the demonstration of the task, because it has already been shown once. Additionally, in all conditions it is gazed elsewhere mostly in p1 and p2 and not during the transportation of the cups in a1, a2 and a3.

5. Conclusion

To conclude, we did nd ostensive signals in tutoring situations in adult-robot interaction. On the one hand, our results for range and velocity show signi cantly exaggerated

Variable	ACI		AAI		ARI		ACI vs AAI	ACI vs ARI	AAI vs ARI
	M	*SD*	*M*	*SD*	*M*	*SD*	*Z*	*Z*	*Z*
velocity a1	3.58	0.81	4.72	1.39	2.08	0.86	−2.394**	−3.668***	−3.747***
velocity a2	4.19	1.84	6.39	1.71	2.59	0.87	−2.535**	−2.792**	−3.982***
velocity a3	6.62	2.43	11.78	2.95	3.73	1.51	−3.098***	−2.956**	−3.982***
range a1	4.22	2.49	3.41	0.72	6.29	5.53	−0.211	−1.369+	−1.288+
range a2	2.19	0.48	1.88	0.25	2.72	0.97	−1.549+	−1.314+	−2.635**
range a3	1.57	0.37	1.35	0.09	2	0.56	−1.479+	−2.409**	−3.396***
total length eye-gaze to i.p. in	10.86	14.52	6.65	7.15	6.65	7.15	−0.833	−1.419+	−0.76
total length eye-gaze to i.p. a1	27.81	25.02	9.01	16.92	9.25	11.38	−2.2*	−1.882*	−0.97
total length eye-gaze to i.p. p1	24.19	28.17	3.7	9.71	7.35	8.78	−1.853*	−1.03	−1.634+
total length eye-gaze to i.p. a2	15.39	16.67	2.42	4.44	3.16	4.81	−2.054*	−2.066*	−0.244
total length eye-gaze to i.p. p2	33.73	24.63	2.61	7.09	2.69	5.9	−3.055***	−3.306***	−0.082
total length eye-gaze to i.p. a3	23.05	23.09	4.37	8.71	6.2	10.48	−2.273*	−2.292*	−0.384
total length eye-gaze to i.p. su	43.8	23.81	27.55	7.43	19.66	13.65	−0.493	−2.793**	−1.878+
total length eye-gaze to o. in	69.29	29.43	82.32	22.47	62.65	8.7	−1.353+	−1.15	−2.817**
total length eye-gaze to o. a1	70.94	22.72	89.52	16.69	83.21	13.46	−2.1*	−1.213	−1.155
total length eye-gaze to o. p1	60.95	26.97	88.99	23.87	68.36	25.95	−2.273*	−0.714	−2.097*
total length eye-gaze to o. a2	82.68	18.18	96.2	8.19	92.43	7.85	−2.198*	−1.308+	−1.533+
total length eye-gaze to o. p2	65.02	25.55	97.39	7.09	80.23	22.36	−3.055***	−1.503+	−2.092*
total length eye-gaze to o. a3	76.95	23.25	95.63	8.71	87.23	13.77	−2.273*	−1.252	−1.721*
total length eye-gaze to o. su	55.79	22.63	52.71	31.88	57.92	17.94	−0.352	−0.109	−0.527
total length eye-gaze e. in	20.89	29.12	11.03	18.15	34.93	9	−0.624	−1.984*	−3.127***
total length eye-gaze e. a1	1.91	4.75	1.48	4.67	7.53	10.61	−0.52	−1.625+	−1.919*
total length eye-gaze e. p1	16.09	19.93	7.32	23.14	24.29	26.94	−1.501+	−0.812	−1.952*
total length eye-gaze e. a2	2.51	3.9	1.37	4.34	4.41	7.42	−1.178	−0.371	−1.604+
total length eye-gaze e. p2	2.38	5.35	0	0	17.08	20.59	−1.382+	−1.879*	−2.551**
total length eye-gaze e. a3	0.74	1.67	0	0	6.57	12.94	−1.382+	−0.877	−1.803*
total length eye-gaze e. su	1.09	2.31	7.65	11.74	22.42	15.92	−1.091	−3.507***	−2.267*

Table 1. Results of Mean, Standard deviation, Mann-Whitney U test, $+p$ <0.1, $*p$ <0.05, $**p$ <0.01, $***p$ <0.001, interaction partner ($i.p.$), object ($o.$)., else ($e.$). su = sum = summary, in = intro = introduction

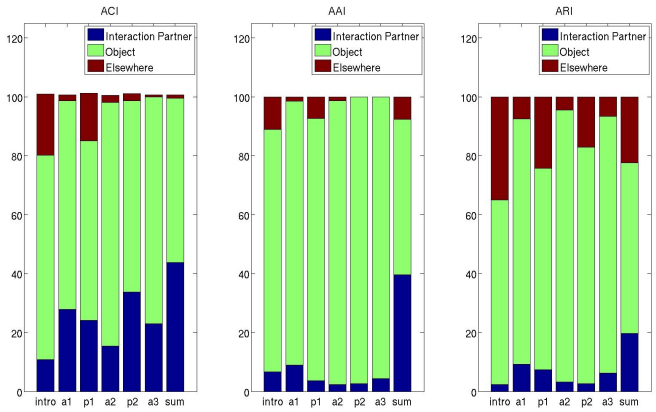


Figure 6. This graph shows the total length of eye-gaze bouts to the interaction partner, the object and somewhere else (y-axis) over time: all seven action parts are displayed (x-axis) for ACI (left), AAI (middle) and ARI (right) condition.

hand movements which are clearly distinguishable from those observable in adult-adult interactions and which are even more accentuated than the hand movements in child-directed tutoring. Thus, ostensive stimuli are present in robot tutoring. These however change over time as we have seen: range of motion decreases drastically, whereas velocity increases slowly. We therefore hypothesize that the reason for this lies in the behavior of the learner which shapes the behavior of the tutor as stated for eye gaze behavior and hand movements by Pitsch et al. [11]. This process could be interpreted as an alignment process where the tutor starts of by clearly signaling his intention of tutoring the infant. This signal decreases during the ongoing interaction while the tutor captures the infant's attention and while observing an understanding process in the infant. The final behavior may thus be described as consisting of fragmentary cues rather than the complete and exaggerated signal. On the other hand, our results reveal that in order to create a contingent interaction with the partner, the learner needs to produce a suitable feedback. This means that although the tutor's hand movements in robot-directed tutoring seem to be even slower and less round than in child-directed tutoring, the tutor's eyegazing behavior in robot-directed tutoring is suggestive of a lack of appropriate social signals on

the recipient's side: The percentage of time the interaction partner is viewed by the tutor is much lower in ARI than in ACI.

The ostensive signals considered here appear practical for the robot to detect situations in which it is being tutored, but we argue that a robot cannot make use of an important ostensive stimulus such as contingency without providing the right signals for the interactional construct. In detail, we nd that already from the introduction on: the eye-gaze behavior in the ARI situation is rather similar to that of the AAI situation, with less time of the eye-gaze being spent on the interaction partner. This is congruent with previous ndings from [14]. If we hypothesize that eye-gaze is also being used in order to check for understanding of the partner, the eye-gaze behavior directly after the end of a subaction becomes relevant. Indeed, we can see that the eye-gaze lengths in both pauses p1 and p2 are signi cantly longer in ACI as opposed to AAI. Thus, the parents appear to look for understanding in their infants. Interestingly, the behavior in ARI tends to be similar to the one in AAI indicating that adults behave differently towards robots. However, in p1 we see a trend for the eye-gaze lengths to be signi cantly longer in ARI as opposed to AAI. This might indicate that the subjects are looking out for signs of understanding in the robot as well. Yet, this behavior dramatically changes in p2 where the eye-gaze length is again decreased to the level of AAI, whereas it is even slightly increased in ACI. This may be interpreted as a reaction to missing signals of understanding from the robot. In the summary part of the action (sum), nally, the overall eye-gaze length towards the robot becomes signi cantly shorter than in both, ACI and AAI.

In order to con rm these results and our interpretation we are planning to carry out analyses of the joint eye-gaze behavior. We hypothesize that the robot is not able to establish mutual gaze especially in the pauses which then leads to the increase of eye-gaze towards the robot.

6. Outlook

These ndings suggest that ostensive signals are present in human-robot tutoring situations and may be used for the robot to learn. However, in order for the robot to elicit a contingent interaction, it needs to provide ostensive signals that indicate its understanding. Based on our observations of the infants' behavior, these ostensive signals have to pertain to attention. That is, the robot has to provide eye gaze that signals attention and establishes joint attention as well as shared attention. Another behavior of the infants that was not modeled in the ARI condition was their attempts to reach and grasp the demonstrated objects. Further analyses need to be carried out in order to reveal the pattern of these reaching gestures - rst impressions of the data suggest that they are far from random but only appear at the end of the demonstrated actions. If this is true, the reaching gestures could be interpreted as a signal that the infant has understood the goal of the action, or at least, the end of the action. Further signals which can be observed from the infants are facial expressions. Again, systematic analyses need to be carried out, but rst impressions suggest that emotional feedback indicates affective reactions to the objects themselves, but also to the attention grabbing behavior of the tutor, and the reaching of the goal.

References

[1] http://www.mangold-international.com/en/products/interact.html.

[2] R. Brand, D. Baldwin, and L. Ashburn. Evidence for'motionese': modi cations in mothers' infant-directed action. *Developmental Science*, 5(1):72 83, 2002.

[3] R. Brand, W. Shallcross, M. Sabatos, and K. Massie. Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action. *INFANCY*, 11(2):203 214, 2007.

[4] G. Csibra and G. Gergely. Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance*, 21, 2005.

[5] A. Fernald and C. Mazzie. Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27(2):209 21, 1991.

[6] L. Gogate, L. Bahrick, and J. Watson. A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4):878 894, 2000.

[7] J. Iverson, O. Capirci, E. Longobardi, and M. Cristina Caselli. Gesturing in mother-child interactions. *Cognitive Development*, 14(1):57 75, 1999.

[8] F. Loemker, 2007. `http://icewing.sourceforge.net`.

[9] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, volume 81, pages 674 679, 1981.

[10] Y. Nagai and K. Rohl ng. Can motionese tell infants and robots what to imitate?. In *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*, pages 299 306, 2007.

[11] K. Pitsch, A. Vollmer, J. Fritsch, B. Wrede, K. Rohl ng, and G. Sagerer. On the loop of action modi cation and the recipients gaze in adult-child-interaction.

[12] K. Rohl ng, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183 1199, 2006.

[13] A. Senju and G. Csibra. Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9):668 671, 2008.

[14] A. Vollmer, K. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. Rohl ng, and B. Wrede. People modify their tutoring behavior in robot-directed interaction for action learning. In *Proceedings of the International Conference on Development and Learning*, 2009.

Canal9: A Database of Political Debates for Analysis of Social Interactions

A.Vinciarelli, A.Dielmann, S.Favre and H.Salamin,
Idiap Research Institute, CP592 - 1920 Martigny (Switzerland)
Ecole Polytechnique Fédérale de Lausanne (EPFL), 1050 Lausanne (Switzerland)
{vincia,adielman,sfavre,hsalamin}@idiap.ch

Abstract

Automatic analysis of social interactions attracts major attention in the computing community, but relatively few benchmarks are available to researchers active in the domain. This paper presents a new, publicly available, corpus of political debates including not only raw data, but a rich set of socially relevant annotations such as turn-taking (who speaks when and how much), agreement and disagreement between participants, and role played by people involved in each debate. The collection includes 70 debates for a total of 43 hours and 10 minutes of material.

1. Introduction

As automatic analysis of social interactions attracts increasingly more attention in the computing community [3] [8], publicly available benchmarks become a crucial element for the progress of the domain. Benchmarks allow different researchers to apply the same experimental protocols over the same data and this is the only way to perform rigorous comparisons between results achieved by different researchers and using different techniques.

This paper presents a corpus of political debates allowing the analysis of important social phenomena like roles (functional and social), conflicts, dominance, agreement and disagreement, status display, communication effectiveness, personality, persuasion, etc. From a social interaction analysis point of view, political debates represent an excellent resource for two important reasons:

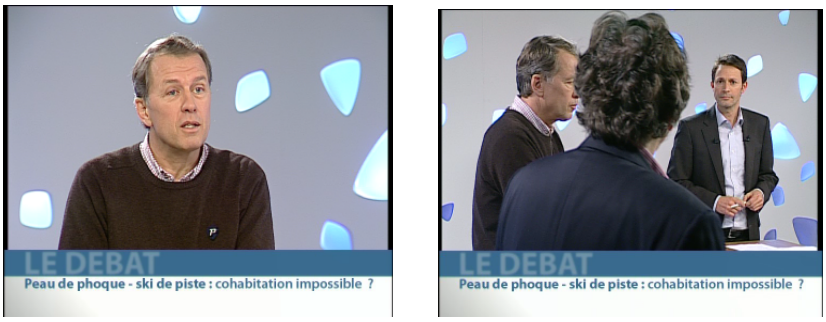
- ***Realism***. In contrast with most benchmarks (for example [1] [5]), political debates are real-world data. Debate participants do not *act* in a simulated social context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections). Thus, even if the debate format imposes some constraints, the participants are moved by real motivations leading to highly spontaneous social behavior.
- ***Privacy issues***. Social interaction recordings are collected, in general, applying the *Informed Consent* principle [2]: subjects must know that they are recorded and must have the right of destroying, partially or totally, the data where they are portrayed. The result is that the subjects tend to be less spontaneous and to eliminate data showing attitudes they do not consider appropriate. As debates are public events, participants know that they are recorded (the principle is respected), but at the same time they are encouraged to be fully spontaneous because this is the only way to be successful in the debates. Furthermore, they cannot destroy the data because these are typically broadcasted live.

The corpus presented in this work includes 70 recordings for a total of 43 hours and 10 minutes of material. Each debate revolves around a *yes/no* question like "*Are you favorable to new laws on scientific research?*". The participants state their answer (*yes* or *no*) at the beginning of the debate and do not change it during the discussion. Each debate involves a moderator that tries to give the same space to all participants (or at least to the two fronts corresponding to *yes* and *no* supporters). Furthermore, the moderator tends to reduce tensions when the discussion becomes too heated.

While including a rich set of annotations, the current version of the dataset is only a first release that will be further enriched in the years to come. Indeed, the Canal9 database is currently used in the core activities of the *Social Signal Processing Network* (SSPNet), a European Network of Excellence aimed at studying Social Signal Processing, and further socially relevant annotations will be added in the framework of this project. The database (including the annotations) will be made publicly available through the web-portal of the SSPNet, at `http://www.sspnet.eu`. The data will be available to any academic and research institution upon signature of a End User Licence Agreement (EULA).

The rest of this article describes the data in terms of media format (Section 2), group composition (Section 3), duration statistics (Section 4), and available annotations (Sec-

978-1-4244-4799-2/09/$25.00



Full Group Personal Shot Multiple Participants

Figure 1. Most frequent camera views.

tion 5).

2. Format and Structure

The recordings are available as high-quality full-frame (720×576 pixels) DV compressed PAL recordings, along with an uncompressed audio stream sampled at $48\ kHz$. They have been *live edited* and, in contrast with corpora collected in laboratory settings, not all the participants are visible all the time. All debates took place in the same recording studio (with no audience) and Figure 1 shows some of the most frequent camera views: *full group* (19.7% of data time), *personal shots* (66.1% of data time), and *multiple participants* (11.0% of data time). The remaining 3.2% corresponds to short reports (typically at the beginning of the debate) and credits shown at both beginning and end of each debate[1].

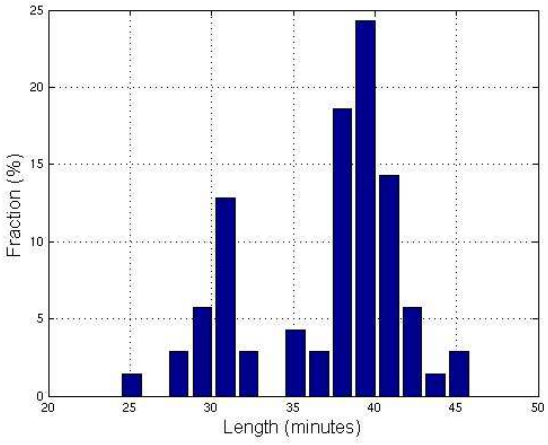


Figure 2. Length Distribution.

3. Group Composition

One of the most important aspects in any group of interacting individuals is the *composition*, that is number and type of people involved [4]. Political debates include two main roles: *moderator* and *participant*.

3.1. The Participants

Each debate revolves around a central question with a *yes/no* answer like "*Are you favorable to the new laws on scientific research?*". Debate participants state explicitly their answer (*yes* or *no*) at the beginning of the discussion and this determines two factions expected to oppose one another during the entire discussion. The spatial arrangement of the participants reflects this situation (see *full group* view in Figure 1). The two factions physically oppose one another in a spatial arrangement that has been shown to elicit agreement between people on the same side and disagreement between people on opposite sides [6]. Overall there are 190 unique participants, 154 participate only in one debate, 25 participate in two debates, and the remaining 11 participate in three. In terms of gender, the set of the participants includes 25 women and 165 men.

[1]The statistics have been extracted from a sample of 10 randomly selected debates.

3.2. The Moderator

All debates include one moderator expected to ensure that all participants have at disposition the same amount of time for expressing their opinion. Furthermore, the moderator intervenes whenever the debate becomes too heated and people tend to interrupt one another or to talk together. Overall, there are five different moderators, 1 woman and 4 men. The woman moderates 28 debates, while the men moderate 24, 9, 8 and 1 debates, respectively.

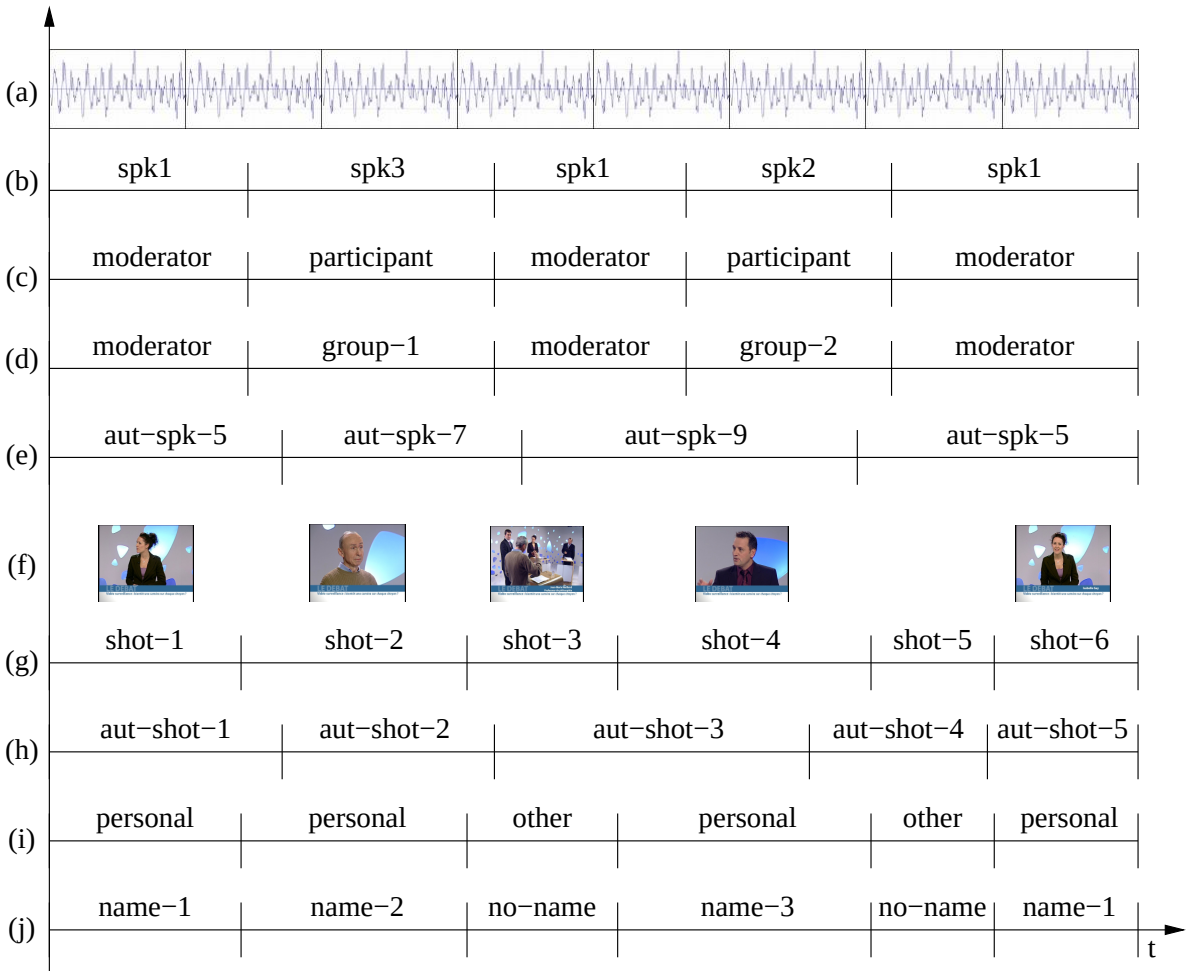


Figure 3. Annotations. The figure shows how the audio channel (a) is annotated in terms of manual speaker segmentation (b), role (c), agreement and disagreement (d), automatic speaker segmentation (e). Furthermore, the figure shows how the video channel (f) is annotated in terms of manual shot segmentation (g), automatic shot segmentation (h), manual shot classification (i), manual identification of participants in personal shots (j).

4. Duration Distribution

In total, the 70 debates of the corpus correspond to 43 hours, 10 minutes and 48 seconds. Of these, 41 hours 50 minutes and 40 seconds (96.9% of the total) correspond to actual discussions, while the remaining time includes reports and credits shown at beginning and end of each debate. The duration changes at each debate and the corresponding distribution is available in Figure 2.

5. Annotations

The political debates are coredated with a wide spectrum of annotations:

- ***Manual Speaker Segmentation***. The audio of each debate (see Figure 3a) has been manually segmented into single speaker intervals (see Figure 3b). Speakers are identified with a label that does not correspond to their names, and all the turns (single speaker segments) where the same person talks hold the same label. The segmentations are stored as `trs` files, an XML format used by the publicly available *transcriber* annotation tool[2].

- ***Role***. The annotations report the *role* played by each person involved in the debates (see Figure 3c), i.e. *moderator* (the journalist expected to guarantee that all persons have enough time to express their opinion and that tries to inhibit aggressive and impolite behaviors) or *participant* (the persons that support one of the two answers to the question around which the debate revolves).

[2]Available at `trans.sourceforge.net/en/presentation.php`.

- ***Agreement and Disagreement***. The participants (see point *Role*) are labeled in terms of *group-1* and *group-2* according to how they answer to the central question of the debate (see Figure 3d). Participants belonging to the same group agree with one another, while participants belonging to different groups disagree with one another.

- ***Automatic Speaker Segmentation***. The output of an automatic speaker diarization system (see Figure 3e) is available for the audio channel of each debate. This allows one to perform experiments where the speker segmentation is supposed to be performed automatically. Furthermore, the availability of both manual and automatic speaker segmentations allows one to estimate the effect of speaker segmentation errors. The segmentations are available as `trs` files (see *Manual Speaker Segmentation* point).

- ***Manual Shot Segmentation***. The video channel of each debate (see Figure 3f) is manually segmented into shots (see Figure 3g), i.e. time intervals between two changes of camera. The shot segmentation is available as a list of shot boundaries, i.e. time instants where the camera changes. The boundaries are stored in ASCII files.

- ***Automatic Shot Segmentation***. The output of an automatic shot segmentation system is available for the video channel of each debate (see Figure 3h). This allows one to perform experiments where the shot segmentation is expected to be performed automatically. The availability of both manual and automatic shot segmentations allows one to assess the effect of shot segmentation errors. The format of the automatic shot segmentations is the same as the one of the manual ones.

- ***Manual Shot Classification***. Each shot is annotated in terms of two classes (see Figure 3i): *personal shot* (see Figure 1) and *other*. This allows one to identify those segments that are particularly suitable for behavior analysis as they clearly show a single person. No automatic classification is available.

- ***Manual Identification of Participants in Personal Shots***. All personal shots showing a given *participant* are annotated with her/his identity (see Figure 3j). This allows one to select only those personal shots where a given participant appears. No automatic version of this annotation is available.

6. Conclusions

This paper has described the first release of the *Canal9* collection of political debates, a corpus aimed at the analysis of social phenomena taking place in competitive discussions. The corpus includes more than 40 hours of videos fully annotated in terms of a rich set of socially relevant features (turn-taking, agreement-disagreement, role) as well as low level descriptors (speaker segmentation, shot segmentation, identity of people appearing in personal shots, shot classes).

The corpus is publicly available through the web-portal of the *Social Signal Processing Network* (`www.sspnet.eu`) upon signature of an appropriate End User Licence Agreement. In its present form, the collection has been used in at least two works recently published in the literature [7] [9]. Further relesases will be available in the next years and will include benchmarking procedures allowing rigorous comparisons of different results.

Acknowledgments. This work has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet). The authors wish to thank Canal9 for kindly allowing the diffusion of the data.

References

[1] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

[2] R. Faden, T. Beauchamp, and N. King. *A History and Theory of Informed Consent*. Oxford University Press, 1986.

[3] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, to appear*, 2009.

[4] J. Levine and R. Moreland. Small groups. In D. Gilbert and G. Lindzey, editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press, 1998.

[5] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri. A multimodal annotated corpus of consensus decision making meetings. *The Journal of Language Resources and Evaluation*, 41(3-4):409–429, 2008.

[6] N. Russo. Connotation of seating arrangements. *The Cornell Journal of Social Relations*, 2:37–44, 1967.

[7] A. Vinciarelli. Capturing order in social interactions. *IEEE Signal Processing Magazine, to appear*, 2009.

[8] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing, to appear*, 2009.

[9] A. Vinciarelli, H. Salamin, and M. Pantic. Social signal processing: Understanding social interaction through nonverbal behavior analysis. In *Proceedings of the International Workshop on Computer Vision and Pattern Recognition for Human Behavior*, 2009.

An Automatic Approach to Virtual Living based on Environmental Sound Cues

[1]Mostafa Al Masum Shaikh, [1]Antonio Rui Ferreira Rebordao, [2]Arturo Nakasone, [2]Helmut Prendinger and [1]Keikichi Hirose

[1]Department of Information and Communication Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan
{almasum, antonio, hirose}@gavo.t.u-tokyo.ac.jp
[2]Digital Contents and Media Sciences Research Division, National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430 Tokyo, Japan
{arturonakasone, helmut}@nii.ac.jp

Abstract

This paper presents a novel indoor and outdoor monitoring system based on sound cues that can be used for the automatic creation of a Life-Log, health care monitoring and/or ambient communication with virtual worlds. Basically, the system detects daily life activities (e.g., laughing, talking, traveling, cooking, sleeping, etc.) and situational references (e.g., inside a train, at a park, at home, at school, etc.) by processing environmental sounds, creates a Life-Log and recreates those activities into a virtual-world. It is easily extensible, portable, feasible to implement and reveals advantages and originality compared with other life-sensing systems. The results of the perceptual tests are encouraging and the system performed satisfactorily in a noisy environment, attracting the attention and curiosity of the subjects.

1. Introduction

Speech is regarded as the most meaningful acoustic event but other types of sounds also convey meaningful information. In a typical environment the human activity is characterized by a multitude of sounds, either produced by humans or by their interaction with objects and devices. Consequently the processing and identification of these acoustic events can be of primordial importance to describe the human and social activities that take place in a certain environment. For example, the jingling sound of cooking utensils (like cooking pan, spoon, knife, etc.) may lead to infer that someone is cooking, the sound of vehicles passing may suggest that someone is on the road, the sound of people speaking mixed with the sound of cutleries may lead us to think that people are in a restaurant, and so on. Such context awareness based on environmental sound cues is the focus of Acoustic Event Detection (AED) that is a recent sub-area of Computational Auditory Scene Analysis (CASA). AED processes acoustic signals and converts them into symbolic descriptions corresponding to a listener's perception of the different events included in those acoustic signals. Several approaches to sense the environment are available [1, 2, 3] and, in the advent of an ubiquitous society, we perceive here a huge potential to combine Virtual Living with context awareness based on environmental sensing. Virtual living is the concept of living and interacting in a virtual-world where each person is represented by an avatar. Thus a user can recreate real-world activities into the virtual-world and interact with other users through avatars. For example, in some contexts (e.g. health care), Virtual Living could be applied to monitor the user's well-being or behavioral abnormalities by someone (e.g. relatives, nurses or care-givers). By nourishing such vision in mind we apply the AED's concept to automatically create a Life-Log that can be represented in the virtual world. We envision that in the future, with the proliferation of the computing power of hand held devices (HHD), availability of internet connectivity and improvements in communication technologies such ambient communication to the virtual world will be common practice in our daily life and will allow us to create vivid and intelligent online social networks. To clarify our motivation let's consider the following scenario: Sami, Anny, Harry and Silvia became friends in a virtual-world but in real life they live in different parts of the world. They often login to a virtual-world and frequently update their status to let others know what they are doing. They use Second Life (SL) [4] to interact with each other in the virtual-world. They are looking forward to use a HHD (e.g. iPhone) that can automate the process of updating their status. Let's assume that they have installed our system in a HDD. In that case the system can capture and process the environmental sounds with a certain interval of time. The processed sound cues can be used together with common sense knowledge to infer the present activity and automatically reflect that activity into the virtual-world. For example, if Silvia is cooking in real-world (i.e., the sound cues are like cutting onions on a chopping board, water falling on a sink, cooking pan and arranging plates), her friends see her moving around the kitchen in SL performing similar activities.

In this paper, we compare the activity's recognition performance of the system with that of human subjects. The second main concern of this paper deals with the automatic generation of virtual-world activities. Since we are dealing with highly varying acoustic sources where practically any imaginable sounds can occur, we have limited our scope in terms of location and activities to be

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

recognized to a particular location (a kitchen). The paper is organized as follows: Section 2 reviews the background studies related to this research. Our approach, in terms of system architecture and description of the system components is explained in section 3. Section 4 refers to the experimental setup, tests and evaluation. Conclusions are presented in Section 5.

2. Background

A number of researchers have investigated the inference of Activities of Daily Living (ADL). In [5], the authors have successfully used cameras and a bracelet to infer hand washing. The authors of [6] used audio-frequency-identification (RFID) tags functionally as contact switches to infer when the users took their medication. The system discussed in [7] used contact switches, temperature switches and pressure sensors to infer meal preparation. The authors of [8] used cameras to infer meal preparation and in [9], the authors used motion and contact sensors, combined with a custom-built medication pad to get a rough inference of meal preparation, toileting and medication consumption. A custom wearable computer with accelerometers, temperature sensors and conductivity sensors to infer activity level is used in [10]. The author of [11] used 13 sensors to infer energy usage in a house, focusing on the use of the heating system. Motion detectors to infer rough location were used in [12] and several sensors like motion sensors, pressure pads, door latch sensors and toilet flush sensors were used in the system described in [13]. The authors of [1] have monitored bathroom activities based on sounds and the system referred in [2] utilized RFID tags to detect objects and thereby infers activities based on the interaction with the detected objects. The research on MIT's house project [14] places a single type of object-based adhesive sensors that provide data that later can be used for kitchen design, context sampling and ADL monitoring. All of these systems perform high-level inference by coarse sensor data and their analysis, and some have added special pieces of hardware to improve their performance, but even so, progress towards accurate ADL detection is still far from desirable. Furthermore, very few researchers reported results of any preliminary user testing [5, 9, 12, 13]. The level of inference using sensors has often been limited, for example, to acknowledge that a person entered the living room and had spent time there. Research that aims at detecting hand washing or tooth brushing almost do not have common synergy, each of them using its own set of idiosyncratic sensors and algorithms. Furthermore a home deployment kit designed to support all these ADLs would be a mass of incompatible and non-communicative widgets. Our approach does not need all these paraphernalia and it's focused on a general inference engine that infers activities from the sound cues that abound in many environments. A similar approach to automatic virtual living is automatic Life-Logging. The idea of a Life-Log can be traced back at least 60 years [15]. Since then a variety of modern projects have spawned such as *the Remembrance Agent* [16], *the Familiar* [17, 18], *myLifeBits* [19], *Memories for Life* [20] and *What Was I Thinking* [21]. In [22] the authors evaluated the user's context in real time and then used variables like current location, activity and social interaction to predict moments of interest.

A Life-Log includes people's experiences that are collected from several sensors and stored in a mass storage device. It is used to support user's memory and satisfy needs for personal information. If someone wants to inform other people about his experiences, he can easily share them by providing his Life-Log. However, a user cannot automatically mirror/reflect his current movements, activities or surrounding environment (e.g., park, shopping mall, etc.) to the virtual life of his avatar. Only very recently [23], the mapping of real-world activities to virtual worlds has been attempted by processing data collected from multiple sensors along with inference logic for real-world activities, but inferring human activity using such data is often inaccurate and insufficient. Furthermore, deploying a sophisticated ubiquitous sensor network in an outdoor environment is expensive and difficult to implement. Our work differs from others in four key ways:

1. instead sensors or video cameras, we use microphones and environmental sound cues to infer location and interaction with objects;
2. due to its portability and simplicity of usage, by using a microphone to capture environmental sounds, it is possible to monitor outdoor environments (e.g. the road, a park, a train station etc.) that previous research almost could not perform;
3. in our model it is easy to incorporate a new set of activities for further needs by just adding more appropriate annotated sound clips and re-training the Hidden Markov Model (HMM) based recognizer;
4. the system can be used as a Life-Log or bridging someone's real-world and his virtual world.

Second Life is a 3D virtual world developed by the Linden Lab in 2003 that allows its users to interact with each other (through avatars or agents) and even trade virtual properties and services [4]. SL's interface is based on a SL client. Its objects are created using the Linden Scripting and in 2007 an alternative product called Open Simulator [24] was released, that aims at developing open source server software for SL clients.

MPML3D is the acronym of Multimodal Presentation Markup Language 3D and it is a XML-scripting language that describes the behavior of agents controlled by computer [25]. These agents are virtual characters (e.g. an avatar in SL) that can emulate human behavior. These presentations are described by the MPML3D script file and currently are supported by SL and Open Simulator. Basically, MPML3D can be used to manipulate avatars that perform presentations in SL or in Open Simulator (these presentations can include sounds, gestures and movement). In order to be able to communicate and perform presentations with the virtual-world using

MPML3D, it's necessary to use the MPML3D Front-End and, in some cases, a Speech Cube and/or Gesture Pack. The MPML3D Front-End receives instructions and communicates them to SL, where they will be executed into animations, scenes and gestures. The virtual agents are SL avatars and, by default, these avatars do not have voices and/or certain gestures capabilities. Thus, in order for them to speak and perform some specific animations, they need to possess the Speech Cube and/or the Gesture Pack that is available in [26]. A MPML3D Back-End and a MPML3D Front-End are available for download at [27]. For more information please refer to [26].

```
<MPML3D version="1.0">
 <Head>
  <Entities>
    <Entity type="human" name="avatar1" resourcePath="girl">
     <Property name="agent_name">Your Second Life avatar's
name here</Property>
        </Entity>
  </Entities>
 </Head>

 <Body startImmediately="task1">
  <Task name="task1" priority="0">
   <Action>avatar1.write("Hello World!" </Action>
   </Task>
 </Body>
</MPML3D>
```

Figure 1: An example of a MPML3D script

3. The System

The goal of our system is to detect daily life activities (e.g., laughing, talking, traveling, cooking, sleeping, etc.) and situational references (e.g., inside a train, at a park, at home, at school, etc.) by processing environmental sounds, creating a Life-Log and recreating those activities into a virtual-world. For example, while the system identifies cooking pan's jingling and chopping board sound as consecutive cues and if the system's local time indicates evening, then from common-sense database the system infers this activity as "cooking" and an avatar would perform this activity in the virtual-world.

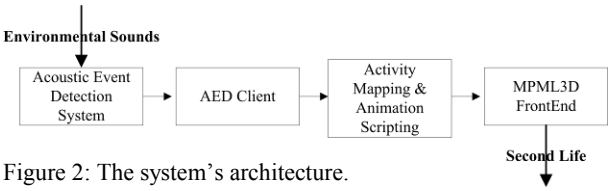


Figure 2: The system's architecture.

The system is based on a pipeline structure and its components are described as follows:

3.1. Activity Event Detection System

Because of their ubiquity we plan to use hand held devices (e.g., a portable computer or a smart phone) to deploy this application. According to the system's architecture (as it can be seen in Figure 3), environmental sound signals are processed and the input is recognized as a set of object labels by a HMM based label recognizer. The detected object list, some common-sense knowledge regarding human activity, object interaction and temporal information (e.g., morning, noon, etc.) are used by the inference agent to infer the activity and the location of the user. Then, the activity and the location are mapped to the virtual world by a scripting language. In the following sections the system's components are described.

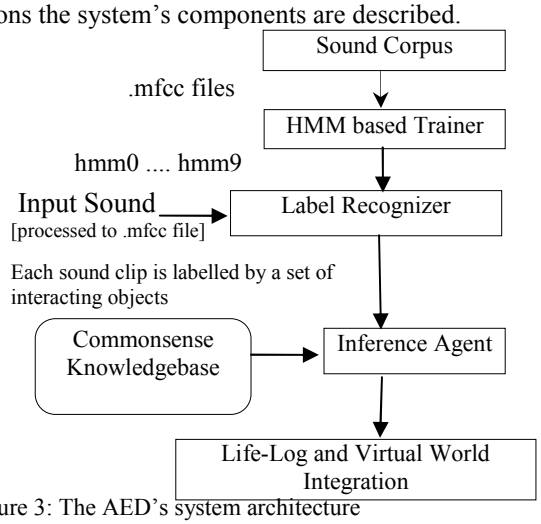


Figure 3: The AED's system architecture

3.1.1 Sound Corpus

Sound patterns are a function of many environmental variables like size and layout or the indoor environment, material of the floors and walls, type of objects (e.g., electrical or mechanical), background noise, etc. The system corpus was designed having in mind the environments where it is going to be used. An analogous way is followed in speech recognition where the system is trained individually on each user for speaker dependent recognition. Therefore, we collected sound samples from different outdoor and indoor places in Tokyo. For a clear audio-temporal delineation during the system training, capturing the sound for each activity was carried out separately. Several subjects were used to collect the corpus sounds. We used a digital SANYO sound recorder (model number: ICR-PS380RM) and the signals were recorded as Stereo, 44.1 KHz, .wav formatted files. Background sounds were recorded simultaneously and the variability of the captured sounds for each activity provides a realistic input for the system training, and increases the robustness and predictive power of the resultant classifier. Some sounds (e.g., water falling, vacuum cleaning machine sounds, etc.) are generally loud and fairly consistent but, on the other side, hands washing, drinking and eating exhibited a high degree of variability. We collected 114 types of sounds and each sound type has 15 samples with length varying from 10 to 25 seconds.

3.1.2 HMM-based trainer

Sound Clip Annotation: We listed 63 objects to assess the objects' interaction in a given sound sample. An annotator opened a sample sound file with WaveSurfer and set the annotation configuration as "HTK Transcription", selected a particular portion of a sound sample and annotated it accordingly to a list of 63 objects. Sometimes a sound sample contains a mixture of sounds produced by different objects. Then, if several sounds

location	Activities
Living Room	Listening Music, Watching TV, Talking, Sitting Idle, Cleaning (vacuum-cleaning)
Work Place	Sitting idle, Working with PC, Drinking
Kitchen	Cleaning, Drinking, Eating, Cooking
Toilet	Washing, Urinating
Gym	Exercising
Train Station	Waiting for Train
Inside Train	Travelling by Train
Public Place	Shopping, Travelling on Road
On the Road	Traveling on Road

Table 1: A list of locations and activities of interest to our study.

overlap, the annotator would be allowed to tag a maximum of 2 objects to annotate such complex sound.

Training Features: A simple frequency characterization would not be robust enough to produce good classification results. To find representative features, a previous study [28] carried out a comparative study on various transformation schemes, including Fourier Transform (FT), Homomorphic Cepstral Coefficients (HCC), Short Time Fourier Transform (STFT), Fast Wavelet Transform (FWT), Continuous Wavelet Transform (CWT) and Mel-Frequency Cepstral Coefficient (MFCC). It was concluded that MFCC might be the best transformation for non-speech environmental sound recognition. A similar opinion can be also found at [29, 30]. These findings motivated us to use MFCC for extracting the features for the sound classification.

The input signal is first pre-emphasized with an Impulse Response filter 1-0.97z^(-1) and MFCC analysis is performed in 25 ms windowed frames advanced every 10 ms. A 39 coefficient vector is extracted from each signal frame window and for each one, the following coefficients are extracted as a feature vector:

- The 12 first MFCC coefficients [c1,…, c12];
- The "null" MFCC coefficient c0, which is proportional to the total energy in the frame;
- 13 "Delta coefficients", estimating the first order derivative of [c0, c1,…, c12];
- 13 "Acceleration coefficients", estimating the second order derivative of [c0, c1,…, c12].

3.1.3 Label Recognizer

Training is performed in the training set (recordings and their associated object labels). The HMM parameters are iteratively optimized with the Baum-Welch algorithm to find a local maximum of the maximum likelihood objective function. We modeled (using the HTK Toolkit [31]) each sound using a left-to-right 88-state (63 for simple object tag + 25 for complex object tag). Each HMM state was composed of two Gaussian mixture components. After that a model initialization stage was done, all the HMM models were trained in eight iterative cycles. For classification, continuous HMM recognition is used and the grammar was chosen in a way that there is no predefined sequence for the activities and each label may be repeated many times in any sequence.

3.1.4 Common-sense Knowledgebase

Once we get the list of objects involved in the recognized sound samples, we must define the object probabilities with respect to the activities of our interest. Requiring humans to specify these probabilities is time consuming and, instead that, the system utilizes a technique adopted from Semantic Orientation (SO) [32, 33] employing the NEAR search operator of the web-search results of the AltaVista Search Engine.

List of objects, $O = \{O_1, O_2, \ldots, O_K\}$ (K=63)
List of locations, $L = \{L_1, L_2, \ldots, L_M\}$ (M=9)
List of activities, $A = \{A_1, A_2, \ldots, A_N\}$ (N=17)
$WL_i = \{WL_1, WL_2, \ldots, WL_P\}$. For example, L_1= "kitchen" and it is represented by $W_{kitchen}$ = {"kitchen", "cookhouse", "canteen", "cuisine"}
$SA(O_i|L_j)$ = Semantic Associative value representing the object O_i to be associated with location L_j
$SA(O_i|A_j)$ = Semantic Associative value representing the object O_i to be associated with activity A_j
The formulas to get the *SA* values are,

$$SA(O_i \mid L_j) = \log_2\left(\frac{\prod_{W \in WL_j} hits(O_i \text{ NEAR } W)}{\prod_{W \in WL_j} \log_2(hits(W))}\right) \quad (1)$$

$$SA(O_i \mid A_j) = \log_2\left(\frac{hits(O_i \text{ NEAR } A_j)}{\log_2(hits(A_j))}\right) \quad (2)$$

3.1.5 Inference Agent

The system continuously listens to the environment and records sounds for 2.5 seconds with an interval of 2.5 seconds between two recordings. Then an object-mapping module provides a list of objects pertaining to the recognized sound classes and the system gathers a list of objects that is compared with the Semantic Associative (SA) value of the activities and with the locations stored in the common-sense knowledgebase.

3.2. AED Client

The AED client connects to the AED server, receives its output and sends this information to the Activity Mapping and Animation Generator block. All the system components were implemented in Java [34] and the system was designed to support several clients.

3.3. Activity Mapping and Animation Generator

This module is the responsible for logging the avatars, mapping the real-world activities with the activities in the virtual-world and defining the presentation. The mapping is pre-defined by the developers in an xml file and, for each activity received as input, a certain sequence of animations, gestures and sounds will be activated and sent to the MPML3D FrontEnd.

3.4. MPML3D FrontEnd

The MPML3D FrontEnd transforms the sequence of desired gestures, movements and sounds into MPML3D scripts that can be interpreted and executed by SL.

4. Test, evaluation and Online Demo

Perceptual tests were carried out, the system was implemented and tested in a noisy real-world environment and an Online Demo was made to illustrate our concept and system.

4.1. Tests and Evaluation

The system was trained and tested to recognize the 17 activities referred in table 1. We developed a perceptual testing methodology to evaluate the system's performance under the effect of continuous sound streams. Four hundred twenty test signals were created and each one contained a mixture of three sound clips chosen from the 114 sound types. These 420 test signals are representative sound cues for the 63 objects that represent the 17 activities. Thus, we grouped these 420 test signals into 17 groups according to their expected affinity to a particular activity and location. Ten human (i.e., five male, five female) judges listened to the test signals and inferred the activity from the given list of 17 activities and 9 possible locations (i.e., forced choice judgment). Each judge was given all the 17 groups of signals to listen and assess (each group had 3 to 6 signals of the same sound type). Since human judges judged each signal individually, in order to compare the result with the system, a generalization on the human assessment was done, i.e., a group of signals had at least more than 3 signals and each signal was assigned a location and activity label by the judges. Thus a group of signals obtained a list of locations and activities. We counted the frequencies of the location and activity labels per group assigned by each judge and took the maximum of the respective labels to assign activity and location for that group. Recognition results for activity and location are presented in Figure 4 and 5 respectively.

The recognition accuracy for activity and location is encouraging (66% and 64%, respectively). From Figure 4 and 5, we notice that humans are skillful in recognizing the activity and location from sounds. It is also evident that the system receives the highest accuracy (85% and 81%, respectively) to detect "traveling on road" and being on the "road" respectively, which is a great achievement and pioneer effort for this research because no previous research attempted to infer outdoor activities with sound cues. The correct classification of sounds related to activity "working with pc" and location "work place" performed poorly due to the sounds' shortness in duration and weakness in strength.

The system was developed in Java and tested in a real and noisy challenging scenario (Open House of the Japanese National Institute of Informatics). A small set of activities (greeting, clapping, stirring liquids, chopping vegetables and arranging the cutlery), that could be performed in a kitchen scenario, were carried out by the visitors and the avatar recreated those activities in a virtual kitchen in SL. The system performed robustly and satisfactorily under a very noisy environment, attracting the attention and curiosity of the visitors.

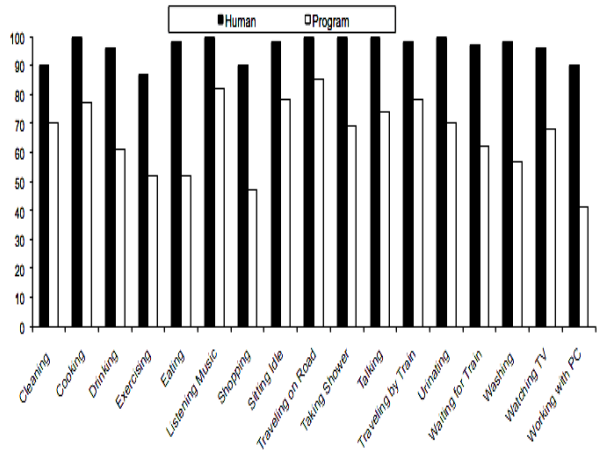


Figure 4: Comparison of the recognition rates for 17 activities, with respect to human judges.

Figure 5: Comparison of the recognition rates for 9 locations, with respect to human judges.

5. Conclusion

This paper describes a novel monitoring system for indoor and outdoor environments. Currently it classifies 17 activities that usually occur in daily life and performs robust activity and location identification by using HMM parameters with MFCC features and a common-sense knowledgebase. The Life-Log of a person can be created automatically and those activities recreated in Second Life. We performed experiments that validate the utility of the system (with an accuracy rate for recognizing activity and location of 66% and 64%, respectively) and tested its robustness in a noisy and challenging real-world environment. The tests performed satisfactorily and we believe that this system contributes positively towards an increased understanding of personal behavioral problems and health monitoring. Soon we plan to test the system in some homes of elderly people living alone in Tokyo. Representing real world activities to a virtual world is surely a source of excitement for the youth but also possesses potential usages related with product or service advertisement, collaborative and e-learning, health care monitoring, etc.

6. References

[1] Kam, A. H., Zhang, J., Liu, N., and Shue, L., 2005. Bathroom Activity Monitoring Based on Sound. In *PERVASIVE'05, 3rd International Conf. on Pervasive Computing.* Germany, LNCS 3468/2005, pp. 47-61.

[2] Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., HahnelInferring, D., 2004. Activities from Interactions with Objects. *IEEE Pervasive Computing*, Vol. 3, No. 4 pp. 50-57.

[3] Temko, A., Nadeu, C., 2005. Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering. In *ICASSP'05,* pp. 505-508.

[4] Linden Research Inc. Second Life. http://secondlife.com/

[5] Mihailidis, A., Fernie, G., and Barbenel, J.C., 2001. The Use of Artificial Intelligence in the Design of an Intelligent Cognitive Orthosis for People with Dementia. *Assistive Technology*, Vol. 13, No. 1, pp. 23–39.

[6] Wan, D., 1999. Magic Medicine Cabinet: A Situated Portal for Consumer Healthcare. In *HUC'99, 1st Int'l Symp. Handheld and Ubiquitous Computing*, LNCS 1707, Springer-Verlag, pp. 352–355

[7] Barger T., Alwan, M., Kell, S., Turner, B., Wood, S., and Naidu, A., 2002. Objective Remote Assessment of Activities of Daily Living: Analysis of Meal Preparation Patterns. *Medical Automation Research Center, Univ. of Virginia Health System.*

[8] Tran, Q., Truong, K., and Mynatt, E., 2001.Cook's Collage: Recovering from Interruptions. *Demo at Ubi-Comp'01, 3rd Int'l Conf. Ubiquitous Computing.*

[9] Glascock A., and Kutzik, D., 2000. Behavioral Telemedicine: A New Approach to the Continuous Nonintrusive Monitoring of Activities of Daily Living, *Telemedicine Journal*, Vol. 6, No. 1, pp. 33–44.

[10] Korhonen, I., Paavilainen, P., and Särelä, A., 2003. Application of Ubiquitous Computing Technologies for Support of Independent Living of the Elderly in Real Life Settings. In *UbiHealth'03, 2nd Int'l Workshop Ubiquitous Computing for Pervasive Healthcare Applications*

[11] Mozer, M., 1998. The Neural Network House: An Environment That Adapts to Its Inhabitants. In *AAAI Spring Symposium, Intelligent Environments, tech. report SS-98-02*, AAAI Press, pp. 110–114.

[12] Campo E., and Chan, M., 2002. Detecting Abnormal Behavior by Real-Time Monitoring of Patients. In *AAAI Workshop Automation as Caregiver*, AAAI Press, pp. 8–12

[13] Guralnik V., and Haigh, K., 2002. Learning Models of Human Behaviour with Sequential Patterns. In *AAAI Workshop Automation as Caregiver*, AAAI Press

[14] MIT house_n Project, http://architecture.mit.edu/house_n

[15] V. Bush, "As we may think", Atlantic Monthly, 1945

[16] B. Rhodes and T. Starner, "Remembrance Agent: A Continuously Running Automated Information Retrieval System," Proc. 1st Int'l Conf. Practical App. of Intelligent Agents and Multi-Agent Technology, 1996, pp. 487-495.

[17] B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," Proc. IEEE Int'l Conf. Acoustics, Speech, and SignalProcessing, IEEE CS Press, vol. 6, 1999, pp. 3037-3040

[18] B. Clarkson, K. Mase, and A. Pentland, "The Familiar: A Living Diary and Companion," Proc. ACM Conf. Computer–Human Interaction, ACM Press, pp. 271-272.

[19] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "MyLifeBits: Fulfilling the Memex Vision," Proc. ACM Multimedia, ACM Press, 2002, pp. 235-238.

[20] A. Fitzgibbon and E. Reiter, "'Memories for Life': Managing Information over a Human Lifetime," UK Computing Research Committee Grand Challenge proposal, 2003.

[21] S. Vemuri and W. Bender, "Next-Generation Personal Memory Aids," BT Technology J., vol. 22, no. 4, 2004.

[22] M. Blum, A. Pentland, G. Troster, et al., "InSense: Internet-Based Life-Logging", IEEE Multimedia vol. 13, Issue 4, pp.40-48, 2006

[23] Mirco, M., Emiliano, M., Nicholas D. L., Shane B. E., Tanzeem, C., Andrew T. C., 2008. The Second Life of a Sensor: Integrating Real-world Experience in Virtual Worlds using Mobile Phones. In Proc. of the Fifth Workshop on Embedded Networked Sensors. Charlottesville, Virginia, June 2-3, 2008.

[24] Open Simulator. http://www.opensimulator.org/wiki/, June 2009.

[25] A. v. Kapri, M. v. Vlie, and S. Ullrich. An introduction to MPML3D. December 2008.

[26] NIIsland. http://slurl.com/secondlife/NIIsland/245/63/25

[27] Global Lab. http://www.prendingerlab.net/globallab

[28] M. Cowling, Non-Speech Environmental Sound Recognition System for Autonomous Surveillance, Ph.D. Thesis, Griffith University, Gold Coast Campus (2004)

[29] H.G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational Auditory Scene Analysis and Its Application to Robot Audition," International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS), (2004), 73–80

[30] A. Eronen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based Context Awareness-Acoustic Modeling and Perceptual Evaluation," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 5, (2003), pp. 529-532

[31] S. Young, The HTK Book, User Manual, Cambridge University Engineering Department, 1995

[32] Hatzivassiloglou, V. and McKeown, K. R., 1997. Predicting the Semantic Orientation of Adjectives. In 35th annual meeting on ACL, pp.174-181

[33] Grefenstette, G., Qu, Y. Evans, D., and Shanahan, J., 2004. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In Computing Attitude and Affect in Text: Theory and Applications, eds. J. Shanahan, Y. Qu, and J. Wiebe, 93-107. The Information Retrieval Series Vol. 20, Netherlands: Springer Verlag

[34] Sun Microsystems. http://java.sun.com/

Social Signals and the action – cognition loop. The case of overhelp and evaluation

Isabella Poggi
Università Roma Tre
Via Manin, 53
poggi@uniroma3.it

Francesca D'Errico
Università Roma Tre
Via Manin, 53
fderrico@uniroma3.it

Abstract

The paper explores the action – cognition loop by investigating the relation between overhelp and evaluation. It presents a study on the helping and overhelping behaviors of teachers with students of their own vs. of a stigmatized culture, and analyses them in terms of a taxonomy of helping behavior, and adopting an annotation scheme to assess the multimodal behavior of teachers and pupils. Results show that overhelping teachers induce more negative evaluations, more often concerning general capacities, and frequently expressed indirectly. This seems to show that the overhelp offered blocks a child's striving for autonomy since it generates a negative evaluation, in particular the belief of an inability of the receiver.

1. Introduction

In Social Signals research, an intriguing subject is the action – cognition loop. In humans social action is mediated by cognition [2]. We decide to do things or to relate to persons on the basis of our conscious or unconscious beliefs; even our emotions, that are a crucial determinant of behaviour, are triggered – either in a reactive or in a reflective way – by beliefs. But at the same time our actions and relations backfire onto our beliefs. A particular case of this action – cognition loop is the relation between help and evaluation. Help is a social action aimed at fulfilling the goals of another person, while evaluation is a set of beliefs concerning how much things, events or persons may favour or thwart our goals. In this work we show how actions and beliefs, help and evaluation may determine each other in a teacher – student relation, and how this link can be assessed by studying their social signals.

2. Evaluation

Evaluating is a cognitive activity of utmost importance in both individual action and social interaction. In every action of our life we form evaluations of objects, persons, events; and at the same time we continuously evaluate other people.

In the goal-and-belief model of mind and social action that we adopt [1, 2, 9], an evaluation is defined as a belief about whether and how much some object, event, person have or provide the power to achieve some goals. The evaluation is positive when they allow and negative when they prevent from achieving a goal, and we evaluate with respect to any kind of goal, utilitarian, ethical, aesthetical; we judge everything as good or bad, ugly or beautiful, useful or unuseful. Evaluating is necessary to action: we evaluate at every moment of our action planning: to decide which goals to pursue, to assess the right actions to do and the tools to use, to check if our goals are achieved.

Yet, beside things or facts we also evaluate persons: we make up an "image" of the persons we meet, that is, a set of evaluative (and non-evaluative) beliefs about them – that person is handsome or ugly, selfish or altruistic, just or unjust, smart or silly; and this determines the social relation we want to have with them. Further, there are two kinds of negative evaluations: one for inadequacy, if someone lacks the power necessary to achieve some goal, and one of noxiousness, if one is actually endowed with power, but a negative power that risks of thwarting someone's goals. A knife is a bad knife if it does not cut well; but it is dangerous if it is too sharp. At school, a teacher may form an evaluation of inadequacy about a boy who is not clever, and an evaluation of noxiousness about one who bothers other children. Moreover, when someone does not succeed in performing some task, one may evaluate negatively either the single performance or – through a process of generalization – the person's general traits or capacities; and the latter is a heavier judgment than the former.

We do not only evaluate others, but also ourselves, thus making up our self-image, a set of evaluative (and non-evaluative) beliefs about ourselves. Our self-image is at least in part determined by our image – how others judge us [8]. But from self image the degree of autonomy of a person depends: if one has a positive evaluation of his own capacities and efficacy, he will pursue his goals in an autonomous and self-confident way. At school, for example, negative evaluations may have a serious impact on a pupil's self image, sense of efficacy, and learning: they tend to dis-able him, to make him less active, and possibly induce him to refrain from action.

3. Help and overhelp

According to the model above, help is a case of goal adoption [2]. An Agent A adopts Agent B's goal when

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

A puts its resources to the service of B's goal, taking it as one's own, and doing actions in order to it. Several types of adoption may be distinguished according to whether they are instrumental to a further goal of the adopter, like in exchange or cooperation, or whether the adopter fulfils the other's goals in a completely disinterested way, like in help and altruism.

Within research on altruistic behaviour, an intriguing issue is the role of helping in social relationships. Help conveys a prosocial intention of the helper, but may also have a negative effect [7], both because the helped person may feel in debt with the helper, and because being helped in itself may perpetuate the dependence of the helped one and possibly the asymmetry of the relationship. This is even clearer with overhelp, that is, when the helper offers his action even if the other could do by himself.

Benevolent overhelp has not been studied in depth so far [7], except for Gilbert and Silvera [6]., who focus on malevolent intentions of the helper to damage the helpee's image in working contexts; furthermore in the few studies on overhelp there's no particular consideration of the helped person . In previous works D'Errico & Leone [3, 4] studied overhelping behaviour in mothers of normal children and of children with a chronic disease, and in teachers with pupils of their own vs. another, stygmatized, culture [5]. In both cases it was found that overhelped children tend to refrain from action, thus failing to achieve autonomy. The reason for this may be that the negative self-evaluation stemming from being overhelped may result in a blow to the image of the helped person, and this in turn may affect her self-image by inducing a lowered aspiration level and a general tendency to de-activation. Further, if this occurs during the learning process, since learning and autonomy are typically made possible by active experience, no activation leads to less learning, more dependence, and less autonomy.

Thus, what evaluations are conveyed, and how, during the learning process, is relevant to predict possible outcomes in the achievement of autonomy.

4. Conveying evaluation

Evaluation and its communication is crucial in social life, and studying the ways in which evaluations are conveyed is a central topic in research on Social Signals.

Evaluations can be communicated by the evaluator, both to the person evaluated and to other persons, either in a direct or an indirect way. Cases of direct evaluation are, for example, praise, criticism or insult, which typically contain an evaluative belief within their very meaning, and typically may affect a person's image. But people care other people's evaluation to such an extent that they may be sensitive to it whatever the channel and the level of explicitness through which they perceive it.

A person may come to believe she is evaluated in some way by someone else in the following ways:

1. direct communication of evaluation, expressed either by verbal or nonverbal communicative signals, e.g. praise, blame, criticism, insult, whether displayed by words, sentences, gestures, grimaces…
2. indirect communication of evaluation. For example, if I tell you this orange is sour, and it is an orange you bought, I may be implying a criticism to your shopping skills
3. bare presupposition of someone's action. If I help you to complete a very easy puzzle, you may infer I think you are not able to do it by yourself.

An important distinction to keep in mind is one between communicated and inferred information. We can define communication [11] as a process in which an Agent S produces a signal in order to a conscious, unconscious or biological goal of having another Agent A come to believe some belief B. On the other hand, inference is a process through which an Agent A, on the basis of some beliefs obtained through perception and/or retrieved from long term memory, and through application of some rules of reasoning, can create a new belief. So it is important to distinguish information that people acquire through communication from one they extract by themselves from the world and from other people's non-communicative behaviour. If I see a person opening his umbrella, I can infer he believes it is raining, even though he did not perform that action *in order to* let me know it's raining. Nonetheless, I can treat that belief just as I treat other information acquired through communication: I can believe it or not, I can myself behave while taking it or not taking it into account – for instance I can decide to open my umbrella too… Other people's assumptions can be understood, used, taken into account irrespective of whether they want to communicate them to us or not, and even whether they themselves are aware or not of their own assumptions.

When a person is helping another, the assumption of an inability of the helped person may "leak" from the helper's behaviour. This assumption may be either indirectly communicated by indirect speech acts (*have you ever made a puzzle?*), or by the direct meaning of nonverbal behaviours (*no, this doesn't go there*), or simply implied by non-communicative actions ("the teacher places the pieces of the puzzle that the pupil could place herself"). It is important to specify that the helper may be in total good faith: she may not be aware at all that a negative evaluation is inadvertently conveyed by her behaviour. Nonetheless, the leaking information may have its effect on the other's image and self-image.

5. The action – cognition loop: overhelp and evaluation

Our hypothesis about the action – cognition loop in helping behaviour is that there is a cognitive mediation between the helper's and the helped person's

behaviour. More specifically, we claim that overhelp may induce an assumption of inability in the helped person and that such assumption may induce less autonomous behaviour and hence, again, a need for help. This may be an undesirable effect, at least in those cases and cultures where individual autonomy is seen as important. In this work we test the first part of the hypothesis – that overhelp induces negative evaluation, by focusing on the behaviours of teachers and pupils in dyadic interactions. In subsequent works, we will test the second part of the hypothesis by showing how pupils tend to act less just in correspondence with teachers' overhelp.

6. An observational study on teachers' help

To investigate the relation between overhelp and evaluation, we based our analysis on a previous study which explored the helping interaction between teachers and their pupils of their own culture or of another, stygmatized, culture [3].

Our study explored whether helping and overhelping behaviour conveys evaluation, and which type it, whether (1) positive or negative evaluation; (2) evaluation on performance or capacity (3) direct or indirect evaluation.

D'Errico et al. [5] carried on a study to analyse the interactions of Italian teachers with their Italian and Rumanian pupils. 21 teacher-pupil dyads of an Italian Primary school (9 with a Rumanian and 12 with an Italian child, all children being between 6 and 8 years old, balanced for gender) were videotaped during a game simulation, designed to possibly imply some crucial helping behaviours of the teacher, but where the teacher could choose either to help the pupil or not.

The Scenario of the game was the Primitive village of the Flintstones family: the pupil played the role of Bam Bam or Pebbles (the Flintstones' little boy and little girl), and the teacher the role of Wilma, the guide who knows all the secrets of the village. After introducing the scenario, the master of the game told the plot and explained that the village was threatened by a magic spell that could be broken by a magic formula. To gain the table with the magic formula the child had to solve a riddle and then, thanks to the solution, could complete the formula by solving a puzzle containing a simple sentence. Both while solving the riddle and completing the puzzle Wilma (the teacher) could choose to help (e.g. simply provide some hints), to overhelp (e.g. tell how to make the complete picture) or not to help at all. Finally, the pupil repeated the magic formula aloud and the master declared the end of the mission because the island was safe.

D'Errico et al. [5] measured how much teachers help Rumanian vs. Italian pupils. In general, data showed that they tend to overhelp, that is, to intrude into the child's autonomous problem solving, more with Rumanian pupils than with Italian ones. Yet, there are large differences between teachers in the amount of overhelp given, and the study distinguished "high intrusive" versus "low intrusive" teachers, depending on the level of overhelp they provided.

In this work we are concerned more on a qualitative than on a quantitative analysis of the teachers' behaviour. Since our hypothesis is that overhelp indirectly conveys a negative evaluation of the helped pupil, we need to assess cases of overhelp and see if evaluative beliefs are contained in the manner it is provided,. So we do not extensively analyse all the teachers of the study, but only two extreme representatives of them: a "high intrusive" and "low intrusive" teacher.

To test our hypothesis that overhelp entails a risk of transmitting a negative evaluation, we need to

- describe and analyse the behaviours of teachers and pupils
- detect which of them imply a goal to help
- quantify the amount of help given, and
- assess whether the teacher's behaviour implies some kind of evaluation, and which one.

To measure the type and amount of help given is a relevant task for research on Social Signals. In fact, since a large part of Social Signals are those that convey information about social relationships, and helping behaviour is a determinant of various social relationships, it is important to have clear in how many ways people can help others. To do so, we built a taxonomy of a teacher's helping behaviours (Sect. 7).

Further, to understand which kind of help is offered in a given interaction, we built an annotation scheme for the analysis of the teacher's behaviour (Sect. 8). A similar scheme will be used in subsequent works to assess the pupil's behaviour).

7. A taxonomy of helping behaviours

According to a view of learning as an active process, teaching can be conceived of as a series of behaviours aimed at providing a person with permanent capacities that make her autonomous, that is, potentially able to solve her own problems, to achieve her goals, by herself. This means that a teacher, when helping a pupil to complete some task, provides adequate help if she takes advantage of task execution to teach him general principles he could eventually transfer to future tasks, while if she does the pupil's job herself, or if she provides help that is not necessary because he could achieve the solution himself, she is overhelping him. More generally, a teacher is overhelping when she definitely tells the pupil what to do, while she is adequately helping when she puts the conditions for the pupil to understand what to do. From this point of view, a teacher's behaviours can be classified as to the extent to which they help the student. Of course, from help to overhelp there is a continuum, but it is possible, in our view, to single out extreme cases.

Table 1 (see below) shows various possible types of helping and overhelping behaviours. Both help and overhelp can be performed through communication, non-communicative action, or finally even by non-action, or better, "deliberate non-action": cases in which a teacher *could* have done something, but apparently *decided not to do* what she could have done. Within all three cases we can distinguish technical, cognitive and affective help or overhelp. The former distinction – communication, action, non-action – refers to the teacher's behaviour, while the latter refers to the processes, in the pupil's mind, to which the teacher's action or non-action is aimed: those which, if favoured by the teacher's intervention, should have an impact over task performance.

Technical help/overhelp is any action or deliberate non-action that directly allows or induces the pupil to perform some moves; cognitive help/overhelp is what provides information or cognitive strategies useful for task completion; affective help/overhelp is what induces affective states that may have an impact over task performance.

Starting from COMMUNICATIVE ACTIONS, typical cases of **technical help** are the communicative actions of providing information, hints, suggestions, but also criticism. Criticising may be seen as a form of adequate help to the extent to which, at least indirectly, it provides positive information as to how to do something. On the other hand, orders, directions, prohibitions can be seen as **technical overhelp**. In fact, we count as overhelp those cases in which the helper is intruding into the helped person's free choice and autonomy. If I tell you: "*there is a nice piece here*", I give you a chance to decide whether or not to place it into the puzzle, while if I tell you "*put this there*", I do not.

Cognitive help includes the communicative actions that do not provide specific solutions but rather reasoning strategies, like when the teacher puts general questions to make the pupil reason, or when she explains processes or proposes doubts while the pupil is making mistakes. Moreover, if a teacher does not only correct the pupil's move, but explains why it is incorrect, making him reflect over his mistaken process of thought, we have a good example of cognitive help. On the other hand, we consider **cognitive overhelp** cases of communication in which the teacher reveals specific moves or strategies the pupil could discover by himself. Again, one may provide both help and overhelp through "affective" communication, that is, communicative acts inducing or preventing emotions that could either favour or hinder the helped person's action. Cases of **helping affective communication** are the communicative actions of encouraging, inciting, praising, confirming, reassuring, sharing emotions with the pupil, and finally minimising his possible negative emotions; while a case of **overhelping affective communication** occurs, for example, if the teacher expresses compassion, or if she hurries the pupils, or simply induces stress in the pupils through leaking of her own anxiety.

Within NON-COMMUNICATIVE ACTIONS, some of the teacher's movements while assisting a pupil are not aimed at communicating but may be nonetheless helping or overhelping actions. Some can be seen as **technical helping behaviours** in that they fulfil the physical conditions for the pupil to do things well: e.g., the teacher preparing the game table, or placing a lamp in the right place to let him see better. But the teacher performs **technical overhelping** through non-communicative actions when she replaces the pupil by making the moves the pupil should do, say, by handing the right piece of the puzzle or placing it herself. She is overhelping also if she undoes his incorrect move, or corrects the pupil's move, by taking away a piece he put into the wrong slot, without telling him why it is wrong. A **cognitive helping non-communicative action** occurs when the teacher does something to put the condition for some cognitive process to take place in the pupil's mind. A typical case is the teacher turning the pieces of the puzzle in the right direction, so the pupil can better see how to place them. In this case, she is not communicating anything, but simply does something that in the pupil might trigger the insight for his problem solving. A **non-communicative cognitive overhelp** occurs if the teacher prevents the pupil from making a mistake, for instance by taking the piece away from his hand, or else if she undoes the pupil's error – say, by removing a piece placed by him – without an explanation. In an active view of learning that aims at developing the learner's autonomy, errors are an important step towards competence. So if the teacher, after the pupil has made a mistake, corrects his move and explains why it is an incorrect move, this is adequate help; but if she prevents him from making errors, or in any other way, she does not give him the chance of understanding why an error is an error, this is overhelp (or, possibly – bad help!). Finally, the teacher's action may fulfil the affective conditions of the pupil's work, by influencing the pupil's emotional state. Thus, it provides **affective help** if it makes the environment warm, motivating or relaxing. Strangely enough, though, it is difficult to find examples of the corresponding affective overhelp in the domain of non-communicative action. If for example the teacher inadvertently expresses her anxiety, thus inducing stress in the pupil, this is a case of communication, albeit unconscious [11]. On the other side, if anxiety simply leads the teacher to do the pupil's moves herself, we see this as technical overhelp, albeit caused by the teacher's emotional state. In such case, her emotion is not communicated but directly *acted out* by performing intruding and overhelping actions. Sometimes a right way to help is **non-action**. Should the teacher hurry the pupil, she might transmit anxiety and make him perform worse: the opposite of this communicative affective overhelp, and sometimes the

best kind of help, is waiting, i.e., refraining from action. Here it is clear how non-action implies a deliberate decision not to act: the teacher is moving her hand toward the puzzle, but then she refrains and puts it behind her hip. This a case of **affective help** through **non-action**. On the contrary, if the teacher stays there doing nothing while the pupil actually would need her help, this is *lack of help*, to be clearly distinguished from deliberate non action. So it is just when you detect movements of inhibition that you can speak of deliberate non-action. Another non intruding way to help are the teacher's **epistemic actions**, i.e., cognitive actions aimed at acquiring knowledge about how the task is being performed. A typical epistemic action is observing the pupil's behaviour attentively to check if he is performing well. Checking and controlling can be defined epistemic actions of acquiring knowledge about how some process is proceeding, in order to be able to re-direct it if something is going wrong. Thus epistemic action may be considered, though indirectly, a case of help, because it is a step before possibly deciding whether to help, and whether to provide technical, cognitive or affective help. Epistemic action may precede, and hence be indirectly, either technical or cognitive or affective help. For instance, if observing the pupil I see he is almost having the insight, but lacks a crucial information, I can provide it, thus giving cognitive help; if I see him discouraged, I can encourage him, providing affective help. On the other hand, the non-action of refraining from doing is most typically a case of affective help, being a way to leave the pupil reflect without hurry or anxiety.

8. An annotation scheme of the teacher's multimodal behaviour

The taxonomy presented so far may help to classify general categories of actions. But to analyze our videos we need to assess the single concrete behaviours performed by teachers and pupils, and classify them as belonging to one or the other of the categories above. So we devised an annotation scheme to analyze teachers' and pupils' multimodal behaviour. The focus of the scheme was to assess the type of social action of the teacher – her possible helping or overhelping behaviour, its possible evaluative import, and its effects over the pupil's reaction.

The annotation scheme is divided into 8 columns (see Tables 1 and 2).

- Column 1 contains the time in the video of the behaviour under analysis.
- In columns 2 and 3, respectively, we describe the teacher's verbal and nonverbal behaviour.
- In col. 4 we write the communicative or non-communicative goal of the behaviours in columns 2 and / or 3. For the verbal behaviour written in col. 2, its goal is by definition a communicative goal, while for the action written in col. 3 the goal to write in col. 4 may be either a communicative goal (for non-verbal communicative signals) or not (for those behaviours that are not intended to provide information).
- Further, since an action – either communicative or not – beside its direct goal may aim at one or more supergoals – other goals for which the direct goal is a means [11] – in col. 5 we write the possible supergoals of the actions in 2 or 3. For a non-communicative action a supergoal is some further effect the agent wants to bring about through goal of col.4: e.g, if a teacher turns the pieces of the puzzle on the right side, she may do so to check the place of the pieces better, and then to know herself where the pieces should go. For a communicative act, the supergoal is an inference the Sender wants the Addressee to draw from that communicative act: if the teacher points at the place in the puzzle where the piece belongs, her communicative supergoal is to suggest the pupil to put it there.
- In col. 6 we classify the goal of col. 4 (or the supergoal written in col 5., when there is one) in terms of the above taxonomy of the teacher's helping or overhelping behaviours (Table 1).
- In col. 7 we write – if there are some – the plausible inferences that one could easily draw from the teacher's actions (columns 2 and 3) and their goals (4 and 5), but that the teacher presumably did not have the goal to be drawn by the Interlocutor. We call them "unwanted inferences" since they are beliefs that may have caused the Agent's action, and since often from the effect we infer the cause, they can be inferred from the Agent's action. Typically, for instance, if the teacher overhelps the pupil by placing some pieces, you may think the teacher doubts s/he is not able to do it.
- In col. 8 we write whether some of the beliefs of columns 4,5 or 7 convey some kind of evaluation, and what kind.

Table 2 (see below) shows three fragments of our analysis. At line 1, time 7.09 (Col.1), the teacher places the two posts of the game in front of the child and orients them toward her (col. 3). Her direct goal, a communicative goal pursued through a nonverbal action, is for the child to pay attention and concentrate to start the game (col.4). This is (col. 6) a Communicative action (C) providing technical help.

Immediately after, at line 2, time 7.10, the teacher bends her head in a head canting posture (col. 3), a posture of welcome, of non-dominance, which means: "I put myself at your level" (4): a Communicative action providing the affective help of making the other feel welcome (6). But this is the typical posture of a mother with her child: a posture of welcome, but marking an asymmetrical relation. So it may have an unintended effect of letting the child infer "your level is low" (7): a negative evaluation (8) inadvertently conveyed by this action.

At line 3., time 7.11, the teacher asks the Rumanian child: *"Have you ever done a puzzle?"* (col. 2). The

direct goal of this question is to check whether the prerequisites are fulfilled for the child to do the game well (col. 4). The reason that may have motivated this check – then the supergoal of this communicative action – is to be certain that the child is not being evaluated (for instance by the experimenter) for a skill in which she has not been trained before (col. 5): then a Communicative Action providing technical help (6). But the presupposition of the question (if I ask you if you have ever done this, I take into account the hypothesis that you have never done this) possibly unmasks an unwanted inference: in your culture such game might not be used as an educational tool (7). This might sound as a possible negative evaluation not simply of the child, but of the whole culture the child comes from (8). Of course, as it can be seen, this kind of analysis leaves room, at least to some extent, to different interpretations. But this is typical of qualitative research, which on the other hand has the advantage of a qualitative and more in depth insight into human cognition and action.

Yet, one might finally decide among different interpretations by taking into account the effects of teachers' different behaviours on pupils. And in fact, as will be clear later, the pupil helped by the high intrusive teacher tended to refrain from action more frequently than the one helped by the low intrusive one.

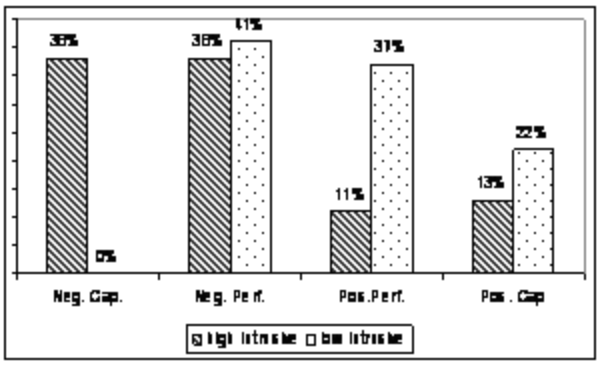
9. Results

As is clear from the taxonomy of helping behaviours presented above, there are many different ways of helping and overhelping. Our study aimed to explore the evaluative effects of help and overhelp in the classroom, trying to focus on the different types of teacher's evaluation that can be directly assumed or indirectly inferred even starting from a benevolent intention to help a pupil.

Since our hypothesis is that help may communicate a feedback about a person's self-image, in the analysis of the teachers' evaluation we distinguished not only positive vs. negative evaluations but also evaluation of performance vs. one of capacity..

A *chi-square* [χ^2 (1, 111)= 33,28; p<0.000] test revealed a significant difference between the low and the high intrusive teacher as to their types of evaluation: as shown in Figure 1., negative evaluation is prevailing in the high intrusive teacher compared to the low intrusive teacher (76% vs 41%); moreover the high intrusive teacher negatively evaluates the children's capacity to solve the problem in 38% of all evaluations. So when she overhelps she sends a negative feedback to the children about himself and his possibilities. The low intrusive teacher negatively evaluates only the child's performance, thus taking care of his self-image, while she provides a good percentage of positive evaluations of the child's capacity (22%).

Figure 1. Types of teachers' evaluation



We attributed progressive scores to the different types of evaluation (1 = negative evaluation of capacity, 2 = negative evaluation of performance, 3 = positive evaluation of performance, 4 = positive evaluation of capacity), to consider a general index of positive evaluation. A t-test shows that the low intrusive teacher generally evaluates significantly in a more positive manner [(t(109):3,951, p<0.000)] as compared to the high intrusive teacher (4.9 vs 8.5).

The results on the different types of evaluation have to be further refined by considering two different ways of communication, direct and indirect.

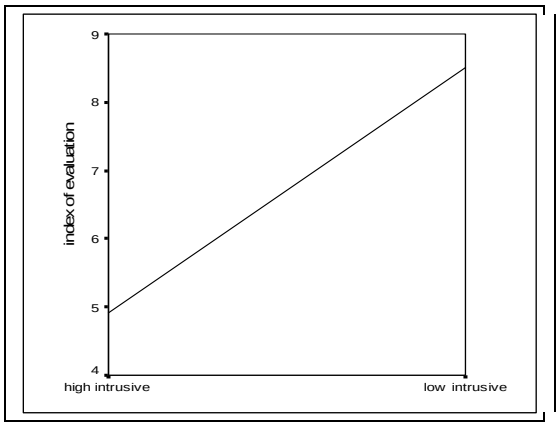
In computing negative evaluations we took into account the ease of the task to be completed by the children, so any kind of intervention that tended to replace e the child in completing the puzzle was labelled as negative evaluation. The high intrusive teacher's evaluations were mainly indirect (67%, vs. 33% direct ones), while the low intrusive teacher used a higher amount of direct evaluations [75% vs 25%; χ^2 (1, 110)= 19,65; p<0.000].

From these results we may conclude that:

- the high intrusive teacher tends to leak more negative evaluations, and more evaluations on the child's capacities than the low intrusive does;
- the low intrusive teacher gives more "constructive" evaluations (that is, more often positive, and less frequently about capacities);
- the low intrusive teacher tends to provide evaluations more in a direct than in an indirect way.

This pattern of evaluative behaviour by the low and high intrusive teachers show that overhelp indirectly lets the pupil infer negative evaluations about him/herself, mainly concerning his/her capacities. This could well account for the subsequent de-activation.found in the previous study. At the same time, that the negative evaluation is mainly indirect especially on the part of the high intrusive teacher might let us think that the more explicitly evaluations are expressed, the better.

Figure 2. Evaluation in high and low intrusive teachers



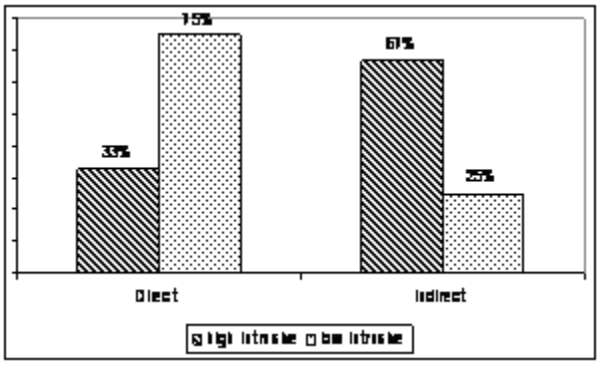


Figure 3. Direct vs. indirect evaluation

10. Conclusion

Evaluation is everywhere. Since evaluating things, events, persons, is of utmost importance in planning to achieve our goals, we tend to evaluate anything and anybody, and exploit any possible chance to evaluate. But evaluation is also food for our image, and to understand how others evaluate us we do not only rely on what they tell us; we try to infer this from their indirect messages, from their action, even from their non-action; and not only when others want to communicate how they value us, but even – possibly, more frequently – if they simply act with us as if they should value that way. We have presented a study on the helping and overhelping behaviour of teachers with pupils from their own and from other cultures, and we have seen how the teacher who overhelps more – the high intrusive one – often leaks a negative evaluation of the child: one that insists on the whole capacity more than on the single performance, and that more often is only indirectly – possibly inadvertently – expressed. This may invest the pupil with a heavier load of negative evaluation, thus possibly leading to refrain from action. In subsequent works we aim to assess the effects of this evaluation load in children's behaviour, to see how much the induction of dependence may be attributed to a blow to image and self-image, thus making it clearer the action – cognition loop.

Acknowledgements

This research is supported by the Seventh Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement Number 231287

References

[1] C. Castelfranchi and I. Poggi. Blushing as a Discourse: Was Darwin Wrong? In R.Crozier (Ed.) *Shyness and Embarrassment. Perspectives from social Psychology*. New: York: Cambridge University Press, pp.230-251, 1990.

[2] R. Conte and C. Castelfranchi. *Cognitive and social action*. London: University College, 1995.

[3] F. D'Errico and G. Leone . Playing to help. Using a game simulation as a tool to observe how mothers of chronic ill children tend to over-help them and how they evaluate their helping behaviours. *Psicologia della salute*, 1: 91-106, 2006.

[4] F. D'Errico Il sovraiuto materno nella malattia cronica infantile: aspetti comportamentali, emozionali e autoriflessivi. In G.Leone (ed) *Le ambivalenze dell'aiuto. Teorie e pratiche del dare e del ricevere.* Milano: Unicopli Editore, 113-172., 2009.

[5] F. D'Errico, G.Leone and T. Mastrovito. The paradox of over-help. When teacher's intervention makes an immigrant child more dependent. In W. Berg (ed) *Multicultural classes* Wiesbaden: Verlag GmbH, in press.

[6] D. Gilbert and D. Silvera, David. Overhelping. *Journal of Personality and Social Psychology*, 70: 678-690, 1996.

[7] G.Leone (ed.) *Le ambivalenze dell'aiuto. Teorie e pratiche del dare e del ricevere.* Milano: Unicopli Editore, in press.

[8] G.H. Mead. *Mind, self and society*. Chicago: University of Chicago Press, 1934.

[9] M. Miceli and C. Castelfranchi, C. The role of evaluation in cognition and social interaction. In K.Dautenhahn (ed.), *Human cognition and agent technology*. Amsterdam: John Benjamins, 1998.

[10] A. Nadler, Relationships, Esteem, and Achievement Perspectives on Autonomous and Dependent Help Seeking, in S.A. Karabenich (Ed.), *Strategic Help Seeking: Implications for Learning and Teaching*, New York, Erlbaum, pp.61-94, 1998.

[11] I. Poggi,. *Mind, hands, face and body. A Goal and belief view of multimodal communication*. Berlin: Weidler, 2007

Table 1.

Teacher's behaviour	Pupil's process	Help	Overhelp
COMMUNICATION	Technical	provides or reminds information, suggestion, hints, soft criticism	orders, directs, forbids
	Cognitive	puts general questions to make the pupil reason and find the solution, explains the process, how one should do, proposes doubts in case of likely mistakes; explains errors	reveals specific moves or strategies
	Affective	encourages, incites, reinforces, confirms, reassures, share and model emotions, minimizes child's negative emotions	expresses compassion, insists in hurrying up, shows anxiety
ACTION	Technical	fulfils technical conditions: prepares game table, put light in the right place	makes pupil's moves substitutes herself for the child
	Cognitive	fulfils cognitive conditions: performs actions to induce insight (turns pieces)	prevent pupil's errors (takes a piece away from the child's hand) or undo pupil's errors (takes pieces put by the child away) without explanation
	Affective	fulfils affective conditions: makes the environment motivating: relaxation, amusement, empowerment, gratification	
NON ACTION	Technical Cognitive Affective	refrain from action: waits, inhibits own action	

Table 2.

1. Time	**2. Speech**	**3. Action**	**4. Goal**	**5. Intended Supergoal**	**6. Type of Action**	**7. Unwanted Inference**	**8. Evaluation**
1 7.09		*Places both posts and orients them toward the child*	Pay attention here and concentrate, let's start the game		C, Technical Help		
2 7.10		*Head canting*	I put myself at your level. I welcome you as a mother with her child		C, Affective Help	Your level is low	Neg, Capacity
3. 7.11	*L'hai mai fatto un puzzle?* Have you ever done a puzzle?		I ask you to confirm if the prerequisites are fulfilled for you to do this game	I do not want you to be evaluated for something no one has taught you	C, Technical Help	May be in your culture these educational tools are not used	Neg, Capacity

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

Social Agents: the rst generations

Dirk Heylen, Mariet Theune, Rieks op den Akker, Anton Nijholt
Human Media Interaction
University of Twente
{heylen,theune,infrieks,anijholt}@cs.utwente.nl

Abstract

Embodied Conversational Agents can be viewed as spoken dialogue systems with a graphical representation of a human body. But the embodiment is not the only difference. Whereas Spoken Dialogue Systems are mostly focused on computing the linguistic dimensions of communication, conversational agents are conceived as intelligent agents that have an identity, a persona. Thus, cognitive modeling is often more involved in ECAs including the modeling of emotion. Whereas spoken dialogue systems are focused on the task, virtual humans are also equipped with social skills involved in interaction. This can take various forms. In this paper we review some of the approaches that have been taken in the first decade of ECA research, by presenting the social signaling skills of three agents we have developed in our group.

1. Introduction

In traditional spoken dialogue systems, the kinds of information services such as TRAINS (http://www.cs.rochester.edu/research/trains/) from the nineties [1], the focus was on getting a speci c task performed by natural language dialogue. The power of a spoken dialogue system is made possible by constraining the domain; which helps semantic processing. Having a clear task, makes it possible to simplify pragmatic processing as well, as the scenario - getting information about a train journey, for instance - is quite well structured, following a simple script. The strategy of such a dialogue system consists in asking a series of questions with restrained options. When the system takes the initiative - starting the conversation with You are talking to the X-system. You can book tickets to destinations from anywhere in Europe. From which city do you want to leave? - this constrains the input suf ciently for speech recognition to perform reasonably well. The spoken dialogue system is thus able to ll in the slots that are needed to formulate a query on its database and provide the user with the information wanted. Besides these information gathering and information providing actions, an important part of the dialogue actions consist in checking whether the system has correctly understood the user - a process referred to in some systems as grounding - and instantiating repair dialogues if this appears not to be the case. A spoken dialogue system is mainly concerned with content and control dimensions of interaction, less with what Goffman has termed the ritual dimension of interaction [10].

To give an idea of the dimensions involved in conversations which have also been found relevant for spoken dialogue systems, Figure 1, shows the major dimensions of conversational activity that are distinguished in the DIT++ taxonomy (http://dit.uvt.nl).

Although the DIT++ scheme provides a slot for social obligation management functions , these are mainly related to formulaic elements of interaction. Spoken dialogue systems such as TRAINS have a similarly limited view on the social aspects. For instance, the choice of wordings may re ect a certain element of formality or politeness and the system will greet the user and present itself.

The virtual humans that have been developed since the start of this century (and perhaps slightly before that), on the other hand, have been endowed with more elaborate social skills. In the following sections we will discuss three embodied conversational agents that have been developed by the Human Media Interaction group, illustrating some of the approaches to equip embodied dialogue systems with social skills. In the nal section we will discuss the approaches and sketch some future developments.

2. The Virtual Receptionist

The virtual agent Karin (Figure 3) was one of the rst embodied dialogue systems that we used to carry out some user experiments. Karin is a virtual receptionist who resides behind the information desk at the lobby of the Virtual Music Center (VMC) - a virtual replica of the theater in Enschede. The dialogue skills of Karin were based on an earlier non-embodied dialogue system that interfaced to a database containing the information on the performances in

978-1-4244-4799-2/09/$25.00

Information transfer functions:
information seeking
information providing functions
Action discussion functions:
commissives and directives
Dialogue Control functions:
feedback
Interaction management:
turn, time and contact management
own/partner communication management
discourse structure management
Social Obligations Management functions:
salutation, self-introduction, apologizing,
gratitude expression, valediction

Figure 1. Dialogue involves several dimensions of activity. Basic information transfer one of these: the dimension of content. Besides information exchange this also involves getting people to do things or committing oneself to do things. On a meta-level, people take actions that ensure that the dialogue goes well: providing feedback on reception of messages or eliciting feedback, for instance. Taking care that the interaction goes well also means that one undertakes actions that make it clear whose turn it is to speak. Besides the level of content/task and control, there is a third level, which is always involved when two or more people interact, the ritual , or social dimension.

```
[K] Hello, I am Karin. How can I help you?
[U] Is there anything on this evening.
[K] There are no performances on the 24th
    of December 1999.
[U] What about the first of January?
[K] On the first of January there is only the
    New Years Concert of Orkest van het Oosten.
[U] Are there other performances in January?
[K] There are 42 performances in the period
    from January 1, 2000 to January 31, 2000.
    Please take a look at the table.
```

Figure 2. Karin: example dialogue

the actual theater. The dialogue system allowed people to query the information about performances and also to order tickets. A short dialogue is presented in Figure 2.

To develop the dialogue system a Wizard of Oz study was conducted in order to get a sense of the kinds of questions people would ask and a sense of how they would ask them. On the basis of this corpus, a parser was developed that uses pattern matching to analyze the user's input. The Karin agent will, as other spoken dialogue system, ask the user questions that will allow it to ll the slots it needs to query the database ([20]).

The introduction of an embodied version of the dialogue system raised questions about the proper way to have the agent behave with its body. What kinds of actions should

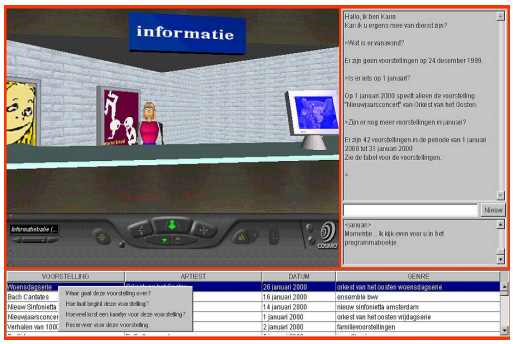


Figure 3. Karin: the virtual receptionist.

it perform? What kinds of nonverbal behaviour should it display and how should this be related to the verbal expressions? In our main study on Karin's nonverbal behaviour we focussed on gaze. Where should the agent be looking at during the course of the interaction?

From the literature on gaze behaviours in interaction, we know that it is involved in several dialogue control function and in interaction management. In a basic sense, gaze is closely related to attention. As a listener, looking at the speaker signals some form of attention which clearly ful ls a contact management role. For a speaker, seeing that the listener is looking, ful lls a typical positive feedback-function. At the end of a turn speakers frequently look to the interlocutor, which can function as an indication that the turn is about to end (turn management). Besides these control functions, gaze can also function as a deictic, pointing device.

The gaze behaviour that we implemented in our agent was related to these conversation regulation aspects and deictic functions. While the user was typing, Karin would look towards the user, as a display of attention. When Karin spoke short sentences she would continue looking at the user, but at the beginning of somewhat longer utterances, we had Karin look away; turning her eyes and head upwards and sidewards. At a certain point she would resume looking at the user. This is similar to the algorithm used in [8]. We also had her look at the table of performances that appears in the screen as a result of a query to direct the user's attention to it.

In an experiment we looked at the effectiveness of this behaviour by comparing three versions of the system. Besides the version that implemented the behaviours mentioned above, we had a version in which Karin looked at the user most of the time and one in which she would change her gaze behaviour in a more or less random way. We had 48 people interact with one of the versions of Karin (16 per condition). They were instructed to make two reservations for a performance. It appeared that subjects who interacted with the system that implemented the gaze algo-

rithm needed signi cantly less time to complete the task. This would indicate that the gaze behaviour had an important part in interaction management, making the conversation go smoother.

Besides keeping track of the time it took the participants to make the reservations, we also asked them to ll out a questionnaire that consisted of several judgements on a ve point Likert scale related to the impression they got from the agent. The factors that we were interested in were *ease of use*, *satisfaction*, *involvement*, *efficiency*, *personality* and the *perceived naturalness of the behaviours*. It is well-known that gaze behaviours also play an important role on the social and affective dimensions of conversations, i.e. gaze plays an important role in social signalling (see [14] for an overview of functions of gaze). It is therefore not surprising that simple differences in the gaze pattern have an effect on the social perception of an agent.

Although we did not nd any signi cant differences between the conditions with respect to judgement of naturalness of eye movements, there were signi cant differences between the conditions on several of the other factors. The version that implemented the algorithm performed the best on the factors *ease of use* (with judgements on statements such as It is easy to get the right information , It took a lot of trouble to order tickets ,...), *personality* (I trust Karin , Karin is a friendly person ,...), and *satisfaction* (I liked talking to Karin , I like ordering tickets this way ,...).

What this indicates is that the nonverbal *behaviours* that may be taken as having primarily an interaction management function also have an effect on the social-affective dimensions. As Goffman already noted, the system (control) functions and the ritual functions cannot be separated, in the sense that whatever behaviour is performed, this may have effects on each of the dimensions[1].

Discussion One should note that the Karin agent, is basically a plain dialogue system with an embodiment added to it. The agent does not have a dedicated reasoning component that deals with the ritual functions of components. The nonverbal gaze behaviours are more or less hard-coded, so to speak, on top of the task-oriented dialogue system. The dialogue system does not provide special variables or modules for personality or friendliness. However, the experiment shows that varying the basic behaviours of an agent has clear effects on how it is being perceived as a social agent.

In the Karin study, users interacted with a real working version of the dialogue system. It showed how certain behaviours have effects on the conversation and the perception of the agent on the social/affective dimension. Agents have been used to learn more about the mapping between social signals and their meanings or effects in other types of studies as well. These may take the form of perception studies, in which subjects are asked to rate the behaviour of an agent on dimensions related to social skills by showing a short video clip. The goal of these studies is to establish some kind of dictionary (or gestionary) of social signals and their meanings. In the context of the SEMAINE project, we have carried out several of such studies ([4], [15], [16], for instance). Although, such studies solve part of the puzzle of associating social signals with their possible meanings, they have several shortcomings. The main problem is that they abstract away the context of the interaction. Showing a video of an agent making a particular gesture, head movement or gaze pattern, does not show the context in which this takes place. In a different context the same signal will often have a different effect as well.

3. The Virtual Tutor

The example of Karin shows that it is practically impossible to dissociate the various dimensions of conversation: content, control and social-emotional factors and that signals for interaction control will also work in part as social signals. In the case of a virtual receptionist, the task as such does not involve very complicated social skills, except perhaps for maintaining some level of politeness. In other kinds of interactions for which virtual agents have been employed, social skills are much more important for the task as such. Consider, for instance, the case of a tutor[2].

A tutor engages in interaction with a student to teach him or her certain knowledge or skills. Typical acts of the tutor include setting speci c objectives for the student, motivating the student, giving instructions, setting a speci c task, asking or answering questions, explaining, providing support, hinting, pumping for more information, giving examples, providing positive or negative feedback and evaluating the student. A tutor does not just need to provide information on an appropriate level in a way that the students can learn optimally, but also has to perform actions that motivate and challenge students. For this, tutors may need to praise or criticize students. A tutor should therefore not just pay attention to how well a student is understanding instructions but also to how the student is feeling.

Lepper ([19]) identi ed four main goals in motivating learners: challenge them, give them con dence, raise their curiosity and make them feel in control. The skills of a good tutor does incorporate social skills. The four motivating goals identi ed by Lepper can be achieved by varying the teaching tactic. Also for a given task, there may be dif-

[1] The interaction of interaction management and social dimensions is also explored in our current work on the perception of different turn-taking behaviours on the perception of the social skills of an agent [24].

[2] In the ECA literature tutors or coaches are popular tasks to study relational aspects of virtual humans ([5], and [12], and [18], are just three early examples), though one of the rst important studies on relational aspects involved a Real-Estate Agent ([7]).

ferent strategies that a tutor can use to reach the learning objective. For instance, the tutor can choose the Socratic method which mainly involves asking questions to the student. This can raise the student's curiosity. This method should be chosen only if the student is quite con dent and has some mastery over the subject. The kind of praise or negative feedback given can provide con dence. The tutor will chose its actions based on how the student feels.

INES is an intelligent tutoring system that was primarily designed to help students practice nursing tasks using a haptic device within a virtual environment ([17]). We paid special attention to affect control in the tutoring dialogues by selecting the appropriate feedback. Also the kind of teaching action, the affective language used, and the overall teaching tactics are adjusted to the presumed mental state of the student. For this, INES takes into account elements of the student's character, his or her con dence level, and an appraisal of the student's actions: did the student make many mistakes, how harmful are the errors that were made, how was the overall performance so far, how active is the student etc. Also taken into account when calculating these values are the dif culty of the task, for instance. This is used to estimated the affective and motivational state of the student (anxious-con dent, dispirited-enthusiastic) as well as the performance on the task.

The tutoring situation is primarily a dialogue, and INES is a combination of an intelligent tutor system and a dialogue manager. The social-affective dimensions affect both the nature of the tutoring and the nature of the dialogue. Affective parameters will affect the style of the feedback. Compare, for instance, It was quite a dif cult task. Try again, but put the needle in more slowly. versus You put the needle in too fast. Try again. This difference in formulations shows the kinds of verbal adaptations the agent is able to make.

Discussion Compared to the Karin agent, INES has modules built in that keep track of the user's mental state and modules that reason about the appropriate action to take, taking this mental state into account. This is re ected in the behaviours that also involve the execution of the task level. In this case different learning strategies may be chosen and actions that differ with respect to presumed con dence. The socio-affective dimension is not only expressed through the choice of learning strategy, but also in the verbal (and to a limited extent nonverbal) expressions that are chosen by the agent. The dialogue acts merge both affective and task dimensions. INES thus shows a different sort of agent compared to Karin, with the social skills intricately mixed in with the task and expressed through strategy and choice of words.

Another important difference relates to the user modeling. In the case of the virtual receptionist, the agent tries to guide the user in providing the information that is needed to make the reservation but is not further concerned with analysing the user's input. INES, on the other hand tries to get a sense of the affective state of the user by interpreting the actions taken and estimating the impact the performance in the exercise might have on the motivational state of the student. Moreover, the INES tutoring agent has an emotional model of its own in which emotional variables such as happy-for or sorry-for are kept track of (for more details see the paper cited).

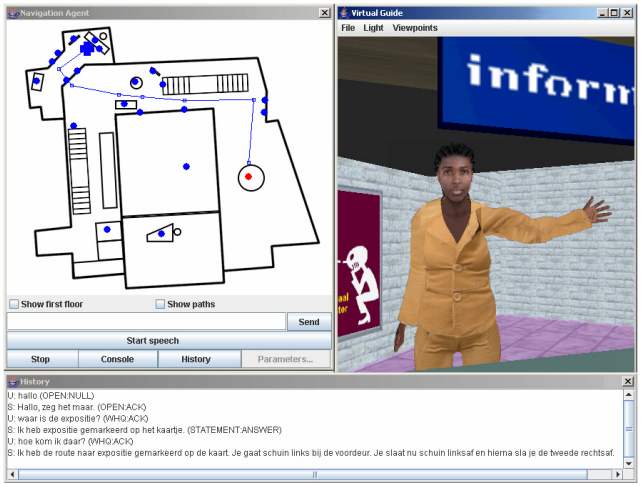


Figure 4. The Virtual Guide

In the next section we present a third virtual human in which social skills are manifested again in a different way. We return to the Virtual Music Center.

4. The Virtual Guide

The Virtual Guide[3] is an embodied conversational agent that also resides in the Virtual Music Center, just as Karin. This agent is able to give directions. Visitors can ask the Guide for information using spoken or typed language as input, in combination with mouse clicks on a map of the environment (see Figure 4). The Virtual Guide responds using spoken language and gestures, and can also show things on the map. In this section we focus on the Guide's verbal behaviour, discussing how the Virtual Guide aligns her level of politeness to that of the user, so as to make her appear more socially intelligent.

Evidence from psycholinguistics has shown that the linguistic representations in social interactions automatically become aligned at many levels [21]. In other words, dialogue partners tend to copy aspects of each other's language. Following Bateman and Paris [3], our notion of alignment includes *affective style*, focusing on the verbal expression of politeness. We have equipped the Virtual

[3]Online demo at http://wwwhome.ewi.utwente.nl/ hofs/dialogue/

Table 1. Some sentence structures that can be handled by the Virtual Guide (translated from Dutch) and their politeness values (P).

Form	Example sentences	P
IMP	Show me the hall.	-3
DECL	You have to tell me where the hall is.	-2
	I have to go to the hall.	-1
	I am looking for the hall.	0
INT	Where is the hall?	0
	Where can I nd the hall?	1
	Would you show me the hall?	2
	Do you know where the hall is?	3

Guide with an adaptive politeness model that dynamically determines the user's level of politeness during the dialogue and lets the Virtual Guide adapt the politeness of her utterances accordingly: a politely worded request for information will result in a polite answer, while a rudely phrased question will result in a less polite reaction.

Like most previous work, we build on Brown and Levinson's politeness theory [6], which is based on the idea that speakers are polite in order to save the hearer's face: a public self-image that every person wants to pursue. The concept of face is divided in *positive face*, the social need for a person to be approved of by others, and *negative face*, the need for autonomy from others. Whenever a speech act goes against either of these needs, this is called a Face Threatening Act (FTA). Brown and Levinson discuss various linguistic strategies to express an FTA at different levels of politeness. The *off-record strategy* is an indirect way of phrasing an FTA so that it allows for a non-face threatening interpretation. For instance, when someone says This weather always makes me thirsty this is probably a hint that he would like a drink. However, for the hearer it is easy to ignore the indirect request and treat the utterance only as an informing act instead.

A dialogue with the Virtual Guide is always initiated by the user, whose rst utterance is then immediately analysed to determine its level of politeness. To this end, we associated the grammar used to parse user utterances with tags indicating their level of politeness on a scale from -5 (least polite) to 5 (most polite). The politeness level depends both on sentence structure, as illustrated in Table 1, and on the use of *modal particles* such as perhaps or possibly , as in Could you perhaps show me the hall? [4] A detailed account of how user politeness is computed can be found in [9]. The system also determines whether the user chooses formal (*u*) or informal pronouns (*je*) to address the Virtual Guide. In its replies, the Guide will use the same choice of pronouns.

[4]Note that the language spoken by the Virtual Guide is Dutch, and the English translations provided in this paper may differ slightly in politeness from their Dutch counterparts.

After having analysed the user's utterance, the Virtual Guide determines the affective style of its reaction. Its degree of alignment to the user can be changed, with the guide adapting its style immediately or only over a series of interchanges.

The rst step in output generation is the selection of a sentence template with the desired level of politeness, computed from the politeness of the preceding user utterance and modi ed by the value of α. Currently the Guide has 21 different politeness tactics at its disposal, including those from Table 1; for a full overview see [9]. The tactics are grouped in clusters of sentence templates with an associated politeness range (e.g, from 4 to 5). During generation, the Virtual Guide randomly selects a template from the appropriate range. This way, a tting template is guaranteed to be found, and some output variation is achieved even when politeness stays at the same level during the dialogue. Finally, gaps in the templates are lled in with formal or informal second person pronouns depending on the user's pronoun choice.

We evaluated the politeness model using both interactive experiments and quantitative evaluations where human judges had to rate the politeness level of the verbal strategies of the Virtual Guide. The main quantitative results are that indirect tactics (e.g., Someone should try again) were generally rated as much less polite than predicted. Also, a frequent comment made by our judges was that subjects found more polite phrasings such as If you don't mind out of place in the context of a request to look at the map. They said Why would I mind? , indicating the absence of any threat to autonomy. See [9] for more details.

In a rst interactive experiment, we let 4 naive participants (students from our department, 2 male and 2 female) carry out three dialogues with the Virtual Guide. In dialogue 1, the Guide showed no alignment (α = 1), and in dialogues 2 and 3 the Guide was set at full alignment (α = 0). For dialogue 2 we asked the participants to be polite to the Guide, and for dialogue 3 we asked them to be impolite. They were free to determine the content of the dialogues (while staying within the direction giving domain).

The participants reported that they clearly noticed the effect of alignment in dialogues 2 and 3. Most of them said they liked the Guide's linguistic style adaptation in the polite dialogue 2, but they found it less appropriate in the impolite dialogue 3, due to the nature of the application: it is the Guide's 'job' to provide a service to the user, and the participants felt that in this role the Guide should always be polite, even to impolite 'customers'. Though the users found an impolite guide somewhat inappropriate, they still thought it was 'fun' to see how the Guide adapted its language to theirs, resulting in exchanges such as:

U: How do I get from here to the exposition, pal?

S: I didn't understand what you said, mate.

The participants also commented on speci c politeness tactics used by the Guide. For example, they thought that system utterances such as It looks like I have been able to indicate the exposition on the map , intended to be polite, made the system sound insecure instead. The users also noted that when the Guide was overly polite this could be interpreted as sarcasm. On the other hand, the Guide also sometimes misinterpreted the user's level of politeness. The most striking example is when one user said Help! after the Guide had repeatedly failed to understand him. The system interpreted this utterance as impolite due to the imperative sentence structure, and promptly reacted by also using an imperative: Say it differently.

Discussion Like the virtual tutor, the guide is able to show its social skills through adapting its verbal utterances. The behaviour is changed based on the behaviour of the user and can thus change dynamically. The examples in the user studies point out again, that it is not always easy to associate speci c behaviours with speci c functions. For instance, associating imperative sentences with directness or impoliteness. Content and context remain very important.

Politeness is a social skill that has been studied in several conversational agents. Presumably the rst attempt at implementing politeness strategies was made by Walker et al. [25], with a recent follow-up in [13]. In their approach, the desired level of politeness of an utterance depends on the social distance between the dialogue participants, the power one has over the other, and the estimated face threat posed by the speech act. Other related work is that of [2, 18, 22] on the generation of tutoring responses, also based on Brown and Levinson's theory. All these systems perform politeness generation based on static input parameters, rather than a dynamic user model that is updated during interaction.[5] Aspects that are taken into account in other work but not by our model include social distance and the face threat level of system dialogue acts.

5. General Discussion

In the previous sections we have presented three embodied conversational agents that we have been working on over the course of the last decade. They illustrate a range of ways in which agents can become social interactants. Our aim has not been to provide the full range of possibilities that have been explored in the eld. By way of summary, we would like to point out some major aspects in the design of social agents.

We hope to have made the point clear that conversational agents are not one-dimensional, but are engaged in interactions on different dimensions which we referred to by such names as task and content, control and social-affective. A single behaviour may work on many dimensions in parallel. This is one aspect that makes the mapping between signal/behaviour and meaning/function less straightforward than is sometimes assumed. A better understanding of how signals work together in different conditions is needed but not so easy to achieve. Perception studies tend to decontextualise the signals and offer only limited insight. On the other hand, current video recordings of interactions that are available for analyse are often too particular, or too arti cial. More and better methods and data collections will need to be developed and made available.

Behaviours displayed by conversational agents are unavoidably interpreted by the human interlocutor on multiple dimensions so that agents that are designed for simple dialogue will not escape judgements about their social skills, even though there are no components in the agent that are concerned with social interaction processing. Social skills are not only displayed through nonverbal signals, but also to what is being said and how it is said. Besides that, the way a task is performed may show interpersonal attitudes as well.

The examples we presented in this paper concerned social skills such as displaying friendliness, being able to motivate people and give con dence, and being polite. Other social skills that have been explored in the literature are showing rapport, empathy, or engagement, amongst others (see for instance, [11] and [23]).

The examples have shown that there can be considerable variation in the complexity of modeling social skills. In two of the agents that we presented, some sort of sensitivity to the social-affective state of the human interlocutor has been implemented. Social skills seem to require some understanding of the needs, desires, goals and emotional state of the other, by de nition. Some of the agents that are around have more intricate user models[6] than the agents we have presented. However, in general, the affect and social signal reading capabilities of most agents are rather limited. Not a lot of work on affective computing technology has been integrated in the ECA systems. This is one of the areas where next generations of social agents could improve upon. Undoubtedly, the next generations of social agents will become more versatile in their social skills with new projects dedicated to studying social signalling in human(-machine) interaction.

Acknowledgments This work has been supported in part by the European Community's Seventh Framework

[5] The politeness model proposed by Andre et al. [2] includes the user's emotional state, to be measured using physiological sensors. However, it seems this approach to user modelling has never been implemented.

[6] See some of the conversational agents developed at ICT (http://ict.usc.edu/projects/virtual_humans).

Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet), and in part by the European Community's Seventh Framework programme under agreement no. 231868 (SERA).

References

[1] J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *Proceedings of the 1996 Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 62 70. ACM, 1996.

[2] E. Andre, M. Rehm, W. Minker, and D. Buhler. Endowing spoken language dialogue systems with emotional intelligence. In *Affective Dialogue Systems*, LNCS 3068, pages 178 187, 2004.

[3] J. Bateman and C. Paris. Adaptation to affective factors: architectural impacts for natural language generation and dialogue. In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on User Modeling (UM-05)*, 2005.

[4] E. Bevacqua, D. Heylen, C. Pelachaud, and M. Tellier. Facial feedback signals for ecas. In *In Proceedings of AISB'07: Artificial and Ambient Intelligence*, Newcastle University, Newcastle upon Tyne, UK, April 2007.

[5] T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293 327, 2005.

[6] P. Brown and S. C. Levinson. *Politeness - Some universals in language usage*. Cambridge University Press, 1987.

[7] J. Cassell and T. W. Bickmore. Negotiated collusion: Modeling social languageand its relationship effects in intelligent agents. *User Model. User-Adapt. Interact.*, 13(1-2):89 132, 2003.

[8] J. Cassell, O. Torres, and S. Prevost. Turn taking vs. discourse structure: How best to model multimodal conversation. In Y. Wilks, editor, *Machine Conversations*, pages 143 154. Kluwer, The Hague, 1999.

[9] M. de Jong, M. Theune, and D. Hofs. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 207 214, 2008.

[10] E. Goffman. Replies and responses. *Language in Society*, 5(3):2257 313, 1976.

[11] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *IVA*, pages 125 138, 2007.

[12] J. Grolleman, E. van Dijk, A. Nijholt, and A. van Emst. Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation. In W. IJsselsteijn, Y. de Kort, C. Midden, B. Eggen, and E. van den Hoven, editors, *Proceedings Persuasive 2006. First International Conference on Persuasive Technology for Human Well-being*, volume 3962 of *Lecture Notes in Computer Science*, pages 133 141, Berlin Heidelberg, 2006. Springer Verlag. ISBN=3-540-34291-5, ISSN=0302-9743.

[13] S. Gupta, M. A. Walker, and D. M. Romano. Generating politeness in task based interaction: An evaluation of the effect of linguistic form and culture. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG-07)*, pages 57 64, 2007.

[14] D. Heylen. Head gestures, gaze and the principles of conversational structure. *International journal of Humanoid Robotics*, 3(3):241 267, 2006. ISSN=0219-8436.

[15] D. Heylen. Multimodal backchannel generation for conversational agents. In *Proceedings of the workshop on Multimodal Output Generation (MOG 2007)*, pages 81 92, University of Twente, 2007. CTIT Series.

[16] D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. Searching for prototypical facial feedback signals. In *IVA*, pages 147 153, 2007.

[17] D. Heylen, A. Nijholt, and R. op den Akker. Affect in tutoring dialogues. *Applied Artificial Intelligence*, 1-2(19), 2005.

[18] L. Johnson, P. Rizzo, W. Bosma, M. Ghijsen, and H. van Welbergen. Generating socially appropriate tutorial dialog. In *Affective Dialogue Systems*, LNCS 3068, pages 254 264, 2004.

[19] M. Lepper. Motivational techniques of expert human tutors: Lessons for the design of computer- based tutors. In *In Computers as Cognitive Tools*, page 75105. Lawrence Erlbaum Associates, 1993.

[20] A. Nijholt and J. Hulstijn. Multimodal interactions with agents in virtual worlds. In N. Kasabov, editor, *Future Directions for Intelligent Information Systems and Information Science*, volume 45 of *Studies in Fuzziness and Soft Computing*, pages 148 173. Physica-Verlag, Heidelberg, Germany, 2000. ISBN=3-7908-1276-5.

[21] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169 226, 2004.

[22] K. Porayska-Pomsta and C. Mellish. Modelling politeness in natural language generation. In *Proceedings of the Third International Conference on Natural Language Generation (INLG-04)*, LNAI 3123, pages 141 150, 2004.

[23] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artif. Intell.*, 166(1-2):140 164, 2005.

[24] M. ter Maat and D. Heylen. Turn management or impression management? In *Proceedings of 9th International Conference on Intelligent Virtual Agents (IVA)*, Amsterdam, The Netherlands, 2009.

[25] M. Walker, J. Cahn, and S. Whittaker. Linguistic style improvisation for lifelike computer characters. In *Entertainment and AI/A-Life, Papers from the 1996 AAAI Workshop.*, 1996. AAAI Technical Report WS-96-03.

Spotting Agreement and Disagreement: A Survey of Nonverbal Audiovisual Cues and Tools

Konstantinos Bousmalis
Computing
Imperial College
London, UK
k.bousmalis@imperial.ac.uk

Marc Mehu
CISA
Uni. Geneva
Geneva, Switzerland
marc.mehu@unige.ch

Maja Pantic
Computing/EEMCS
Imperial College/Uni. Twente
London, UK/Enschede, NL
m.pantic@imperial.ac.uk

Abstract

While detecting and interpreting temporal patterns of non–verbal behavioral cues in a given context is a natural and often unconscious process for humans, it remains a rather difficult task for computer systems. Nevertheless, it is an important one to achieve if the goal is to realise a naturalistic communication between humans and machines. Machines that are able to sense social attitudes like agreement and disagreement and respond to them in a meaningful way are likely to be welcomed by users due to the more natural, efficient and human–centered interaction they are bound to experience. This paper surveys the nonverbal cues that could be present during agreement and disagreement behavioural displays and lists a number of tools that could be useful in detecting them, as well as a few publicly available databases that could be used to train these tools for analysis of spontaneous, audiovisual instances of agreement and disagreement.

1. Introduction

Agreements and disagreements occur daily in human–human interaction, and are inevitable in a variety of everyday situations. These could be as simple as finding a location to dine and as complex as discussing about notoriously controversial topics, like politics or religion. Agreement and disagreement are frequently expressed verbally, but the nonverbal behavioral cues that occur during these expressions play a crucial role in their interpretation [13]. That is naturally the case not only for agreement and disagreement, but for all facets of human social behavior, including politeness, flirting, social relations, and other social attitudes [78].

Machine analysis of nonverbal behavioral cues (*e.g.* blinks, smiles, nods, crossed arms, etc.), have recently been the focus of intensive research, as surveyed by Pantic *et al.* in [56, 58]. Similarly, significant advances have been made in the area of affect recognition (for exhaustive surveys, see [29, 82]). However, research efforts on the machine analysis of social attitudes are still at a rather early stage [56, 78].

There is no overview available, to the best of our knowledge, of nonverbal behavioral cues exhibited during agreement and disagreement. This paper attempts to fill this gap and to be the first step towards our eventual objective: creating a system that can automatically detect the relevant behavioral cues, and spot agreement or disagreement based on both their presence and temporal dynamics.

In this paper we list *(a)* different nonverbal behavioral cues relevant to detecting agreement and disagreement, *(b)* a number of tools that can detect these cues, and *(c)* a list of databases that can prove useful in the development of an automated system for (dis)agreement detection.

Note that we are interested only in those cues that can be detected using a monocular audiovisual data capture system. The main reason for this choice is the fact that the average user has a monocular camera connected to their computer system and hence, any output from this research will be directly applicable in standard user applications, without the need for additional and expensive equipment (such as biosensors, thermal cameras, *etc.*). Furthermore, it will be possible to directly apply the research findings for automatically analyzing and detecting agreement and disagreement in television data, such as televised political debates.

2. Agreement and Disagreement

Distinguishing between different kinds of agreement and disagreement is difficult, mainly because of the lack of a widely accepted definition of (dis)agreement [13]. Ekman [18] talked about listener's expressions of agreement and disagreement, distinguishing them from the relevant speaker's expressions. Argyle [1] specifically discussed the fact that speakers attend to listeners for nonverbal signals that not only serve as feedback to the process of the conver-

978-1-4244-4799-2/09/$25.00

sation, but also as an expression of the listener's opinion. Seiter et al. [67–69] have specifically discussed the importance of listener's expressions of disagreement.

Based on the findings reported by Poggi [62], we can distinguish among at least three ways one could express (dis)agreement with:

Direct Speaker's (Dis)Agreement: A speaker uses specific words that convey direct (dis)agreement, *e.g.* "I (dis)agree with what you have just said".

Indirect Speaker's (Dis)Agreement: A speaker does not explicitly state his or her (dis)agreement, but expresses an opinion that is congruent (agreement) or contradictory (disagreement) to an opinion that was expressed earlier in the conversation.

Nonverbal Listener's (Dis)Agreement: A listener expresses non-verbally her (dis)agreement to an opinion that was just expressed. This could be via auditory cues like "mm hmm" or visual cues like a head nod or a smile. (For a full list of the nonverbal cues that can be displayed during (dis)agreement, see Tables 1 and 2.)

Moreover, displays of agreement, and especially disagreement, can often be accompanied by expressions of emotions like anger, boredom, disgust or frustration as is the case for disagreement [27, 28, 68]. Hence, if the aim is to develop an automated system for (dis)agreement detection, automatic recognition of these affective states should be a part of the system as well.

In addition, Pomerantz [63, 64] describes disagreement as a dispreferred activity, and states that a weak agreement could actually be a preface to an act of disagreement. This makes the problem of (dis)agreement analysis truly complex. In this paper, we leave this aspect out of discussion.

3. Cues of Agreement and Disagreement

3.1. Cues of Agreement

Table 1 contains a list of all cues that can possibly be present during an agreement act. The most prevalent cue seems to be the **Head Nod** which is believed to be a nearly–universal indication of agreement [14, 50]. **Listener Smiles** are also rather indicative. However, both cues could have different meanings [7, 30], as further explained in Section 3.3.

When it comes to **Eyebrow Raise**, it is believed that it occurs in combination with other agreement–relevant cues particularly during an act of Nonverbal Listener's Agreement [18, 66]. Cohen [13] states that **Laughter** could also increase the reliability of any reasoning about detecting agreement, however there is no statistically grounded work on that, as far as we know. Finally, although **Sideways Leaning**, *e.g.*, leaning on a wall due to relaxation is referred to as an agreement cue by Bull [9] and reiterated by Argyle [1]. However, it is specifically discredited by Bull himself [10] as a weaker sign of agreement.

Human's communication system is fairly complex and it is unlikely that receivers will form intricate representations of attitude on the basis of a single cue. In fact, people most probably infer attitudes like agreement by using a combination of such cues, or through the perception of second order dynamic processes that involve these cues. For example **Mimicry** is a mutual imitation of the interlocutor's nonverbal behaviour and is believed to foster affiliation, agreement, and liking [12]. Mimicking the other person's positive behaviour such as nod or smile could therefore be interpreted as agreement; while the presence of the cue on its own might just signal something else, like submissiveness or interest.

3.2. Cues of Disagreement

When it comes to disagreement, it seems that a head shake is the most common cue. A **Head Shake** could specifically mean the refusal or reluctance to believe what is being said [18]. However, much like the head nod and the smile, this signal can have different purposes (look at Section 3.3 below).

Ironic smiles are a result of a conflict between two set of muscles and therefore are not as naturally occurring as benign smiles [1, 65]. Similar to the ironic smile is the **Cheek Crease**, during which a lip corner is pulled back strongly, deliberately distorting a smile to convey sarcasm [50]. These cues seem to be present in expressions of spontaneous and posed disagreement [50, 68].

Ekman [18] specifies that the **Eyebrow Raise**, or "scowling", as referred to by Seiter et al. [67], may indicate lack of understanding. However, it can also indicate, like the head shake, a listener's inability to believe what the speaker is saying or has just said. It can even express a "mock astonishment", when combined with a raised upper eyelid and/or a jaw drop.

Morris [50] mentions a number of disagreement–related cues. One of them is the **Nose Flare**, a result of the contraction of the muscles on either side of the nose, which is often accompanied by a sharp intake of air. Morris also mentions the **Head Roll**, which is the action of repeatedly tilting the head left and right expressing doubt. The **Sudden 'Cut Off'** is a gaze avoidance in which the head is turned fully away from the speaker. The **Leg Clamp**, though not specifically linked to disagreement, signifies stubbornness, as if the conversation participant was saying: "My ideas, like my body, are clamped firmly in position and will not budge an inch" [50]. The **Forefinger** and **Hand Wag**, during which an erect forefinger or a hand with the palm out-

CUE	KIND	REFERENCES
Head Nod	Head Gesture	[1, 14, 25, 30, 41, 50, 66]
Listener Smile/Lip Corner Pull (AU12, AU13)	Facial Action	[1, 7, 50]
Eyebrow Raise (AU1, AU2) + other agreement cues	Facial Action	[66]
AU1 + AU2 + Head Nod	Facial Action, Head Gesture	[16, 18]
AU1 + AU2 + Smile (AU12, AU13)	Facial Action	[16, 18]
AU1 + AU2 + Agreement Word	Facial Action, Verbal Cue	[16, 18]
Sideways Leaning	Body Posture	[1, 9, 30]
Laughter	Audiovisual Cue	[13]
Mimicry	Second–order Vocal and/or Gestural Cue	[1, 30, 35]

Table 1. Cues of Agreement. For relevant descriptions of AUs, see FACS [19].

wards, respectively, is wagged from side–to–side has a dissenting meaning. The **Neck Clamp**, the **Lip Bite** accompanied by a vigorous head shake, and the **Clenched Fist** signal anger with what is being said. The **Hand Cross** is simply a two–handed version of the hand wag. The **Hand Chop** is the action during which a hand imitates an axe, and the **Hand Scissor** is the action during which the hands imitate the blades of a pair of scissors. Morris mentions that both are often used unconsciously during a heated discussion. **Arm Folding** is widely known as signifying a defensive attitude and could also signify disagreement, *e.g.*, in situations where one participants is being verbally attacked in a strong disagreement [9, 25, 50].

Another very interesting cue is the **Throat Clearing**. Givens [25] states that disagreement and uncertainty can act like chemicals or food irritants and cause this signal. Givens also mentions that **Self–manipulation**, *e.g.*, a finger on the lips, massaging a hand, or a chin rub, can provide self–comfort when politeness prevents a listener from expressing disbelief and disagreement. Moreover, Givens argues that a sudden appearance of **Slightly Parted Lips** is a strong signal of nonverbal listener's disagreement. This is in agreement with Ekman's [18] finding that a listener's preparatory–to–speech mouth movement signals a desire to take the floor. Givens also considers a **Lip Pucker** to be the first sign of disagreement.

Disagreement could also be inferred by second order cues such as interruption, delay in responding, or utterance length. For example, Greatbach et al [28] argued that disagreement can be stronger if an **Interruption** and overlapping speech occur. Similarly, **Delays** in responding could be characteristics of a dispreferred activity, such as a disagreement act [63, 64]. In these two examples, it is not the act of speaking or not speaking *per se* that conveys disagreement but the act of violating implicit rules of turn-taking in a conversation. Note, however, that there are certain cases where disagreement becomes the preferred activity, as is the case with responses to compliments [53]. Finally, **Utterance Length** has been shown to be particularly longer in disagreement than in agreement acts [13, 24].

Table 2 shows a complete list of cues associated with disagreement.

3.3. Backchannel Signals: Nods, shakes and smiles

Ekman [18] states that although emotional expressions during conversations are a reaction to the "affective content", they can also relate to the participants' feelings regarding the nature and progress of the conversation itself, *i.e.*, they can serve as backchannel signals. Brunner [7] specifies that there are three levels of meaning a feedback backchannel could have, with the higher level implying and containing the lower ones. These are: **Level 1**—Involvement, **Level 2**—Level of understanding, **Level 3**—Actual response, *e.g.*, (dis)agreement.

Argyle [1] supports this by stating that backchannel signals may indicate attention and understanding, provide feedback like agreement, or be a part of mimicry, which in turn could signify agreement.

So, agreement and disagreement could be conveyed using backchannel signals and it could be argued that most of the implicit nonverbal cues of (dis)agreement are of this sort. For example, nods and shakes are two of the most common backchannel gestures. Nods usually have an affirmative meaning, especially if they're repeated and their amplitude is large. Smaller, one–way nods usually serve as signals of involvement in the conversation [1, 66]. However, it should be noted that head nods could also be negative [66]. Brunner [7] states that listener smiles can also be backchannels and are used in the same way as head nods. Brunner also argues that smiles act on the third level, *i.e.*, they provide a positive response to what is being said, they provide acknowledgment of understanding, and keep the listener involved in the conversation.

Head shakes are less common, and although they can have a dissenting meaning [1], they could also be part of a question or laughter [1, 30].

CUE	KIND	REFERENCES
Head Shake	Head Gesture	[1, 18, 30, 41, 50, 67, 69]
Head Roll	Head Gesture	[50]
Sudden 'cut off' (of they eye contact)	Head Gesture	[25]
Eye Roll	Facial Action	[41, 67–69]
Ironic Smile/Smirking [AU12 L/R (+AU14)]	Facial Action	[18, 67]
AU1 + AU2 + Raised Upper Lid (AU5)/...	Facial Action	[18]
.../Open Jaw Drop (AU26) with abrupt onset		
Barely noticeable lip–clenching (AU23, AU24)	Facial Action	[25]
Cheek Crease (AU14)	Facial Action	[50]
Lowered Eyebrow/Frowning (AU4)	Facial Action	[25, 69]
Lip Bite (AU32)	Facial Action	[50]
Lip Pucker (AU18)	Facial Action	[25]
Slightly Parted Lips (AU25)	Facial Action	[25]
Mouth Movement (Preparatory for Speech) (AU25/AU26)	Facial Action	[18]
Nose Flare (AU38)	Facial Action	[50]
Nose Twist (AU9 L/R and/or AU10 L/R and/or AU11 L/R)	Facial Action	[50]
Tongue Show (AU19)	Facial Action	[25]
Suddenly Narrowed/Slitted Eyes (fast AU7)	Facial Action	[25]
Arm Folding	Body Posture	[9, 25, 50]
Head/Chin Support on Hand	Body/Head Posture	[9, 25, 50]
Large Body Shift	Body Action	[25]
Leg Clamp (the crossed leg is clamped by the hands)	Body Posture	[50]
Sighing	Auditory Cue	[68]
Throat Clearing	Auditory Cue	[25]
Delays:Delayed Turn Initiation, Pauses, Filled Pauses	Second–order Auditory Cue	[13, 24, 28, 32, 63, 64]
Utterance Length	Second–order Auditory Cue	[13, 24]
Interruption	Second–order Auditory Cue	[28]
Clenched Fist	Hand Action	[25, 50]
Forefinger Raise	Hand Action	[50]
Forefinger Wag	Hand Action	[50]
Hand Chop	Hand Action	[50]
Hand Cross	Hand Action	[50]
Hand Wag	Hand Action	[50]
Hands Scissor	Hand Action	[50]
Neck Clamp	Hand/Head Action	[50]
Self–manipulation	Hand/Facial Action	[25, 50]
Head Scratch	Head/Hand Action	[50]
Gaze Aversion	Gaze	[66]

Table 2. Cues for Disagreement. For relevant descriptions of AUs, see FACS [19]

4. Detection Tools

Although in some cases detecting the cues in Tables 1 and 2 is rather straightforward, as is the case with cues that correspond to Action Units, there are cues that are known to be hard to detect. Two such examples are **Arm Folding** and **Head and Chin Support on a Hand**. [58]

However, there are known techniques that would be able to detect most of the cues listed in Tables 1 and 2. For example, most of the current head pose estimation computer–vision systems (for an exhaustive survey refer to [51]) can be adjusted for detection of **Head Nods and Shakes**, probably the most important cues for our objective. A system that can detect nods and shakes particularly well is the work of Morency *et al*. [47, 48].

There are a few attempts to automatically detect **Mimicry**, one of which is by Meservy *et al*. [45]. Keller *et al*. [35] also mention the possibility of using Motion Energy Analysis [6] to analyze the synchrony between the move-

ments of the participants in a dyadic conversation. Pentland [59] measures mimicry (or "mirroring", as called in [59]) in conversational audio patterns, by using auditory backchannels and short words.

The hand and body actions of **Forefinger Wag**, **Hand Wag**, **Hand Cross** and **Hands Scissor** could be detected with adapted versions of human activity detection methods such as the work of Oikonomopoulos *et al*. [54], Marszałek *et al*. [42], Mikolajczyk *et al*. [46], Laptev *et al*. [37], Niebles *et al*. [52] and Shechtman *et al*. [70]. Actions like **Leg** or **Neck Clamp** and **Arm Folding** could also be detected with adaptions of these methods, but with more difficulty, and both dynamic and static features would have to be used for better results. Motion History Images [6] could also be used for such actions, but they have proven to be particularly sensitive to, *e.g*., different clothing. The latter actions could also be detected by the arm and hand tracker of Buehler *et al*. [8]. Most of the other hand actions, and especially **Hand Chop**, **Hands Scissor**, **Hand Wag and Cross** could also be detected by adapting the latter work. **Clenched Fist** and **Forefinger Raise and Wag** seem to be able to be detected by adapting the hand gesture interface system implemented by Ike *et al*. [33]. Most of the aforementioned hand gestures and some self–manipulation gestures like face/lips touching can be detected by sign language recognition methods such as that by Ding and Martinez [15].

When it comes to automatically detecting facial actions, significant advances have been made over the past ten years. Table 4 lists examples of the state–of–the–art systems, omitting older ones that cannot detect Action Units (AUs) in combinations, as discussed and surveyed by Tian *et al*. [72]. AUs are atomic facial signals, the smallest visually discernible facial movements. FACS [19] defines 9 upper face AUs, 18 lower face AUs, and 5 miscellaneous AUs. The most comprehensive works in automatic AU detection are those of Koelstra and Pantic [36] and Vural *et al*. [79], as they detect most of the AUs defined in FACS [19], including those that could be cues of (dis)agreement. The former also enables analysis of temporal dynamics of AUs, which could prove very important when distinguishing, for example, a smile(slow symmetric action) from a smirk (fast asymmetric action). However, these methods will not work particularly well if rigid head movements are not properly dealt with, which is usually a problem with naturalistic, spontaneous data. The work of Valstar and Pantic [76] can also detect many of the AUs listed in tables 1 and 2, including their temporal dynamics, while handling problems with head movement registration rather well. For exhaustive surveys on the topic, see Pantic *et al*. [55, 58].

Smiles relate to AU12 and AU13, which can be recognized by many AU detection systems, as one can see in Table 4. However, the work done by Valstar *et al*. [75] is able to distinguish between spontaneous and posed smiles, which could prove particularly useful in differentiating between genuine, benign smiles and ironic ones (*e.g*., smirks).

CUE	REFERENCES
Head Nod/Shake	[20, 22, 34, 47, 71]
Mimicry	[35, 45, 59]
Smiles vs Smirks	[75]
Utterance Length	[32]
Laughter	[60, 61, 74]
Eye Roll	[20]
Head Roll	[20]
Filled Pause	[2, 23, 26, 80]
Pause	[4, 43]
Interruption	[38, 40]
Throat Clearing	[44]
Tongue	[21, 83]
Sudden 'Cut Off'	[3]
Hand Scissor/Wag/Cross	[8, 37, 42, 46, 52, 54, 70]
Clenched Fist/Forefinger Raise	[33]
Forefinger Wag	[33, 37, 42, 46, 52, 54, 70]

Table 3. Tools for detecting cues for agreement and disagreement

Sudden 'Cut Off' can be detected by adapting methods aimed at detecting the focus of one's attention such as the recent work of Ba and Odobez [3]. Other works on head tracking [51] and on gaze tracking [49] can be adapted for this purpose as well. Recent work can also detect **Laughter** and distinguish it from speech, using auditory cues [74] or a fusion of auditory and visual cues [60, 61]. Finally, the work of Matos *et al*. in [44] can detect **Throat Clearing** as a sub–goal to cough detection.

Tables 3 and 4 list some of the discussed, recently proposed tools that could be used/adapted to detect the cues relevant to agreement and disagreement, as those listed in Tables 1 and 2. Yet, in spite of this obvious progress in automatic analysis of various behavioural cues, no effort has been reported so far towards automatic analysis of (dis)agreement in naturalistic data. The only work in the field is that by el Kaliouby and Robinson [20], which attempted (dis)agreement classification of acted behavioural displays based on head and facial movements. Detection of these signals in naturalistic data is yet to be attempted.

5. Databases of Relevant Naturalistic Data

To develop and evaluate automatic analyzers capable of dealing with naturalistic occurrences of agreement and disagreement as defined earlier in this paper, large collections of training and test data, recorded in naturalistic settings, are needed.

Televised political debates provide an interesting platform for analyzing agreement and disagreement–related

System	AUs Detected																	
	1	**2**	**4**	**5**	**9**	**10**	**11**	**12**	**13**	**14**	**18**	**19**	**23**	**24**	**25**	**26**	**32**	**38**
Tian *et al*. (2001) [72]	√	√	√	√	√	√		√						√	√			
el Kaliouby *et al*. (2005) [20]	√	√						√	√	√	√				√	√		
Pantic *et al*. (2005) [57]	√	√	√	√		√		√	√		√		√	√	√	√		√
Bartlett *et al*. (2006) [5]	√	√	√	√	√	√	√	√		√			√	√	√	√		
Littlewort *et al*. (2006) [39]	√	√	√	√	√													
Yang *et al*. (2007) [81]	√	√		√		√		√		√								
Valstar *et al*. (2007) [76]	√	√	√	√	√	√		√	√		√			√	√	√		
Koelstra *et al*. (2008) [36]	√	√	√	√	√	√	√	√	√	√	√		√	√	√	√		
Vural *et al*. (2008) [79]	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Tong *et al*. (2009) [73]	√	√	√				√						√	√	√	√		

Table 4. AU detection systems

cues. Since the first televised political debates of the 1960's, debates have become more common, and the audience actually expects the participation of political figures in them. [68] At the same time, the presentation of such debates has evolved from a single–screen approach to multiple split screens, where every reaction each participant makes is available for examination, regardless of who the speaker is. [67] Even if only a single screen is used, the director of the debate will often use close–ups of the speaker or the listeners to give access to the nonverbal aspect of their behavior. [31] Research has suggested that those watching the debates perceive as less likable the participants who attempt to belittle a debate opponent via cues of nonverbal listener's disagreement. Interestingly enough, political figures are still prepped to display certain cues for that purpose, and hence this is an interesting case of acted agreement and disagreement.

Canal9[1] [77] is an example of a database of political debates. The database contains a total of over 42 hours of real televised debates on Canal 9, a Swiss television network. There is always a moderator and two sides that argue, with one or more participants on each side. Although this is a "political" debates database, the subjects are not always politicians, and the public opinion does not matter as much. Hence, instances of masked or acted (dis)agreement mentioned above, are rare. The debates are pre–edited in one feed and more than one camera angles are used.

Roma Tre Political Debates[1] is another such database. It contains ten political talk shows and pre–election debates aired on Italian television networks. The number of participants ranges from two to six and each video lasts from 60 to 90 minutes.

The *Green Persuasive Dataset*[1] is a database of 8 recorded instances of attempts by strong pro–green individuals to convince others to adopt a 'greener' lifestyle. There are many instances of agreement and disagreement. Each discussion is a dyadic interaction and lasts from 25 to 48 minutes.

Other databases that could be useful for training and testing automated tools for (dis)agreement detection would be those capturing the instances of human–human or human–computer interaction, in which occurrences of (dis)agreement are very common. Such databases are group meetings recordings like the *AMI Dataset*[1] [11] and human–virtual character interaction recordings like the *SAL Dataset*[1] [17]. For an exhaustive overview of such databases, see [29, 82].

6. Conclusion

This paper has attempted to provide an overview of the cues useful for detecting agreement and disagreement. It has also attempted to provide a list of the state–of–the–art tools that can be used/adapted to detect these cues. Finally, a list of databases that could be used to train and test automated tools for (dis)agreement detection is also provided. Hence, we hope that the paper can serve as an introductory reading to al researchers interested in the problem of automatic detection of agreement and disagreement.

Acknowledgments

This work has been funded in part by the EC's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 231287 (SSPNet). The work of Konstantinos Bousmalis is also funded in part by the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

[1] These databases and more information about them can be found online at the SSPNet web portal (http://www.sspnet.eu).

References

[1] M. Argyle. *Bodily Communication*. 2 edn., 1988. Chapter 7.

[2] K. Audhkhasi, K. Kandhway, O. Deshmukh, and A. Verma. Formant–based technique for automatic filled-pause detection in spoken english. *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 4857–4860. 2009.

[3] S. Ba and J. Odobez. Recognizing Visual Focus of Attention from Head Pose in Natural Meetings. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 39(1):16–33, 2009.

[4] D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi–party meetings using prosodic and lexical cues. *Proc. Int'l Conf. Spoken Language Processing*, pp. 949–952. 2002.

[5] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 223–230. 2006.

[6] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[7] L. J. Brunner. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728–734, 1979.

[8] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. *Proc. Conf. British Machine Vision*, pp. 1105–1114. 2008.

[9] P. Bull. *Posture and Gesture*, chap. 5: The Encoding of Disagreement and Agreement, pp. 62–69. Pergamon Press, 1987.

[10] P. Bull. *Posture and Gesture*, chap. 6: The Decoding of Interest/Boredom and Disagreement/Agreement, pp. 70–84. Pergamon Press, 1987.

[11] J. Carletta. Unleashing the killer corpus: experiences in creating the multi–everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[12] T. Chartrand and J. Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[13] S. Cohen. A computerized scale for monitoring levels of agreement during a conversation. *University of Pennsylvania Working Papers in Linguistics*, 8(1):57–70, 2003.

[14] C. Darwin. *The expression of emotions in man and animals*. Oxford University Press, USA, 2002.

[15] L. Ding and A. Martinez. Modelling and recognition of the linguistic components in american sign language. *Image and Vision Computing Journal*, 27(12), 2009.

[16] A. Dittmann. Develpmental factors in conversational behavior. *Journal of Communication*, 22(4):404–423, 1972.

[17] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, *et al.* The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:483–500, 2007.

[18] P. Ekman. *Human Ethology*, chap. About Brows: Emotional and Conversational Signals. Cambridge Univ. Press, 1979.

[19] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system. Salt Lake City: Research Nexus, 2002.

[20] R. el Kaliouby and P. Robinson. Real–time inference of complex mental states from facial expressions and head gestures. *Proc. IEEE Int'l Conf. Computer Vision & Pattern Recognition*, vol. 3, p. 154. 2004.

[21] Z. Fu, W. Li, X. Li, F. Li, and Y. Wang. Automatic tongue location and segmentation. *Proc. Int'l Conf. Audio, Language and Image Processing*, pp. 1050–1055. 2008.

[22] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. *Proc. IEEE Int'l Workshop Robot and Human Interactive Communication*, pp. 159–164. 2004.

[23] M. Gabrea and D. O'Shaughnessy. Detection of filled pauses in spontaneous conversational speech. *Proc. Int'l Conf. Spoken Language Processing*, vol. 3, pp. 678–681. 2000.

[24] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. *Proc. Meeting Association for Computational Linguistics*, pp. 669–676. 2004.

[25] D. B. Givens. The nonverbal dictionary of gestures, signs and body language cue. Center for Nonverbal Studies Press, Sokane, WA, 2002.

[26] M. Goto, K. Itou, and S. Hayamizu. A real–time filled pause detection system for spontaneous speech recognition. *Proc. European Conf. Speech Communication and Technology*, pp. 227–230. 1999.

[27] J. Gottman, H. Markman, and C. Notarius. The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior. *Journal of Marriage and the Family*, 39(3):461–477, 1977.

[28] D. Greatbatch. *Talk at Work*, chap. 9: On the management of disagreement between news interviewees. Cambridge University Press, 1992.

[29] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int'l Journal of Synthetic Emotion*, 1(1), 2009.

[30] U. Hadar, T. Steiner, and F. C. Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[31] F. Haumer and W. Donsbach. The rivalry of nonverbal cues on the perception of politicians by television viewers. *Journal of Broadcasting and Electronic Media*, 53(2):262–279, 2009.

[32] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. *Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 34–36. 2003.

[33] T. Ike, N. Kishikawa, and B. Stenger. A Real-Time Hand Gesture Interface Implemented on a Multi-Core Processor. *Proc. IAPR Conf. Machine Vision Applications*, pp. 9–12. 2007.

[34] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the between-eyes. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 40–45. 2000.

[35] E. Keller and W. Tschacher. Prosodic and gestural expression of interactional agreement. *Lecture Notes in Computer Science*, 4775:85–98, 2007.

[36] S. Koelstra and M. Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*. 2008.

[37] I. Laptev and P. Perez. Retrieving actions in movies. *Proc. Int'l Conf. Computer Vision*, pp. 1–8. 2007.

[38] C.-C. Lee, S. Lee, and S. Narayanan. An analysis of multimodal cues of interruption in dyadic spoken interactions. *Proc. European Conf. Speech Communication and Technology*, pp. 1678–1681. 2008.

[39] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[40] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech, and Language Processing*, 14(5):1526–1540, 2006.

[41] V. Manusov and A. R. Trees. "Are You Kidding Me?": The Role of Nonverbal Cues in the Verbal Accounting Process. *The Journal of Communication*, 52(3):640–656, 2002.

[42] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. *Proc. IEEE Conf. Computer Vision & Pattern Recognition*. 2009.

[43] M. Marzinzik and B. Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Trans. Speech and Audio Processing*, 10(2):109–118, 2002.

[44] S. Matos, S. Birring, I. Pavord, and H. Evans. Detection of cough signals in continuous audio recordings using hidden markov models. *IEEE Trans. Biomedical Engineering*, 53(6):1078–1083, 2006.

[45] T. Meservy, M. Jensen, J. Kruse, J. Burgoon, J. Nunamaker, D. Twitchell, G. Tsechpenakis, and D. Metaxas. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, 2005.

[46] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8. 2008.

[47] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 18–24. 2005.

[48] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8-9):568–585, 2007.

[49] C. Morimoto and M. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[50] D. Morris. *Bodytalk: A world guide to gestures*. Jonathan Cape, 1994.

[51] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[52] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8. 2007.

[53] R. Ogden. Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, 38(10):1752–1775, 2006.

[54] A. Oikonomopoulos, M. Pantic, and I. Patras. Sparse B-spline polynomial descriptors for human activity recognition. *Image and Vision Computing*, 27(12), 2009.

[55] M. Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of Royal Society B*, 2009.

[56] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huang. Human–Centred Intelligent Human–Computer Interaction (HCI^2): How far are we from attaining it? *Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[57] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, vol. 4, pp. 3358–3363. 2005.

[58] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang. Human computing and machine understanding of human behavior: A survey. *Lecture Notes in Computer Science*, 4451:47–71, 2007.

[59] A. Pentland. Socially aware, computation and communication. *Computer*, 38(3):33–40, 2005.

[60] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 37–44. 2008.

[61] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. *Proc. ACM Int'l Conf. Content–based Image and Video Retrieval*, pp. 329–338. 2008.

[62] I. Poggi. *Mind, hands, face and body: Goal and belief view of multimodal communication*. Weidler, 2007.

[63] A. M. Pomerantz. *Second Assessments: A Study of Some Features of Agreements/Disagreements*. General sociology, University of California, Irvine, 1975.

[64] A. M. Pomerantz. *Structures of Social Action: Studies in Conversation Analysis*, chap. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. Studies in Emotion and Social Interaction. Cambridge University Press, 1984.

[65] W. Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95(1):52–77, 1984.

[66] H. M. Rosenfeld and M. Hancks. *The Relationship of Verbal and Nonverbal Communication*, chap. The Nonverbal Context of Verbal Listener Responses, pp. 193–206. Walter de Gruyter, 1980.

[67] J. Seiter. Does communicating nonverbal disagreement during an opponent's speech affect the credibility of the debater in the background? *Psychological Reports*, 84:855–861, 1999.

[68] J. S. Seiter, H. J. Kinzer, and H. Weger. Background behavior in live debates: The effects of the implicit ad hominem fallacy. *Communication Reports*, 19(1):57–69, 2006.

[69] J. S. Seiter and H. Weger. Audience perceptions of candidates' appropriateness as a function of nonverbal behaviors displayed during televised political debates. *The Journal of Social Psychology*, 145(2):225–236, 2005.

[70] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8. 2007.

[71] W. Tan and G. Rong. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461 – 466, 2003.

[72] Y.-I. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[73] Y. Tong, W. Liao, and Q. Ji. *Affective Information Processing*, chap. 10: Automatic Facial Action Unit Recognition by Modeling Their Semantic and Dynamic Relationships, pp. 159–180. Springer London, 2009.

[74] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144 – 158, 2007.

[75] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 38–45. 2007.

[76] M. F. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. *Lecture Notes in Computer Science*, 4796:118–127, 2007.

[77] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. *Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interfaces*, vol. 2. 2009.

[78] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12), 2009.

[79] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Automated drowsiness detection for improved driving safety. *Proc. Int'l Conf. Automotive Technologies*. 2008.

[80] C.-H. Wu and G.-L. Yan. Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition. *The Journal of VLSI Signal Processing*, 36(2–3):91–104, 2004.

[81] P. Yang, Q. Liu, and D. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 1–6. 2007.

[82] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[83] L. Zhi, J.-Q. Yan, T. Zhou, and Q.-L. Tang. Tongue Shape Detection Based on B-Spline. *Proc. Int'l Conf. Machine Learning and Cybernetics*, pp. 3829–3832. 2006.

Index