Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Facial Communicative Signals - Valence Recognition in Task-Oriented Human-Robot Interaction

Christian Lang, Sven Wachsmuth, Marc Hanheide, Heiko Wersing

2012

Preprint:

This is an accepted article published in International Journal of Social Robotics. The final authenticated version is available online at: https://doi.org/[DOI not available]

Noname manuscript No.
(will be inserted by the editor)

Facial Communicative Signals

Valence Recognition in Task-Oriented Human-Robot Interaction

Christian Lang · Sven Wachsmuth · Marc Hanheide · Heiko Wersing

Received: date / Accepted: date

Abstract This paper investigates facial communicative signals (head gestures, eye gaze, and facial expressions) as nonverbal feedback in human-robot interaction. Motivated by a discussion of the literature, we suggest scenario-specific investigations due to the complex nature of these signals and present an object-teaching scenario where subjects teach the names of objects to a robot, which shall term these objects correctly afterwards. The robot's verbal answers are to elicit facial communicative signals of its interaction partners. We investigated the human ability to recognize this spontaneous facial feedback and also the performance of two automatic recognition approaches. The first one is a static approach yielding baseline results, whereas the second considers the temporal dynamics and achieved classification rates comparable to the human performance.

Keywords Facial Communicative Signals · Valence Recognition · Head Gestures · Eye Gaze · Facial Expressions · Object Teaching · Active Appearance Models

1 Introduction

The face is a rich means of nonverbal feedback and thus plays an important role in the communication between humans. One important goal of the research on automatic facial signal processing is to enhance human-robot interactions by enabling the robot to react appropriately to the nonverbal feedback of its interaction partner and thus improve the interaction quality as perceived by the user. In this paper, we discuss the usage of facial communicative signals such as head gestures, eye gaze, and facial expressions and their interpretation as nonverbal feedback (Sect. 2). Due to the complex nature of these signals, we advocate a pragmatic simplification and concentration on specific human-robot interaction scenarios. Sect. 3.1 describes such a scenario where subjects teach objects to a robot, which is expected to recognize these objects afterwards. In Sect. 3.2, we report an investigation of the recognition performance of other subjects that interpreted the nonverbal feedback of the robot's interaction partners in terms of valence, i.e. they judged whether the robot termed an object correctly based on the facial behavior of the person who taught the object.

Furthermore, we investigated the automatic recognition of this feedback. A simple static approach using actice appearance models [23] and a SVM classifier provides baseline results and is discussed in Sect. 4.3. Subsequently, a more sophisticated approach that considers the temporal dynamics is presented in Sect. 4.4. It is based on dynamic time warping and the selection of discriminative reference subsequences in the video data and achieved classification rates comparable to the human performance in a subject-dependent classification. Finally, Sect. 5 concludes and remarks on future work.

This work has been supported by the Honda Research Institute Europe, Offenbach, Germany.

Christian Lang
Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, Germany
E-mail: clang@cor-lab.uni-bielefeld.de

Sven Wachsmuth
Applied Informatics, Bielefeld University, Germany

Marc Hanheide
School of Computer Science, University of Lincoln, UK

Heiko Wersing
Honda Research Institute Europe, Offenbach, Germany

2 Facial Communicative Signals

By *Facial Communicative Signals* (FCSs) we mean any visual facial behavior in interaction situations that can be interpreted as nonverbal feedback and can thus be utilized to

infer something meaningful about the course of interaction resp. interaction partner who shows this facial behavior.

This is a pragmatic view. It does not distinguish between deliberately given "signals" and unintentionally present "cues" [81], but focuses on their meaningful interpretability in an interaction situation only. Furthermore, it does not define which facial behaviors exactly are FCSs and which are not; in fact, this might depend on the context of the interaction. However, in general FCSs include head gestures, eye gaze, and facial expressions, which are discussed in the following sections.

2.1 Head Gestures

Birdwhistell [10] reported several different kinds of head nods and sweeps and also a few other head movements and positions that constitute meaningful elements in conversations. Head movements of various velocities and amplitudes occur frequently during speech [54] and are highly synchronized with prosody [48]. Despite large variations depending on several factors (e.g. personality and content of conversation), Graf *et al.* [48] found two different types of nods and one type of swing as typical movement patterns. There are partially large intercultural differences, for example between the indian "head wiggle" and western head gestures of comparable meaning [113].

Heylen [55] compiled a long list of functions head movements can serve in social interactions, including signalling yes/no and various intentions, controlling and organizing the interaction, communicating agreement, support or degree of understanding, stressing certain aspects of the uttering, marking lexical repairs, etc. For instance, Poggi *et al.* [96] investigated in detail the meaning of head nods, Goodwin and Goodwin [47] did this for "thinking faces" when people search for a word, and Hadar *et al.* [54] found correlations between head gestures and speaking resp. listening turns.

2.2 Eye Gaze

In contrast to other primates, the visual appearance of human eyes allow for advanced gaze-signalling and thus enhance communication [67], although other species make some use of gaze in social interactions [36], too. Human brains seem to feature an expert system for gaze perception [101]. People can estimate the gaze direction of others reasonably well, especially when being looked at (e.g. [20]).

The apparently most important single aspect of eye gaze is to signal visual attention (e.g. [72]), it is also very important to establish joint attention (e.g [36]). Evidence suggests that direct gaze of others can automatically draw attentention to them (e.g. [51]) and averted gaze can automatically shift attention away (to possibly important objects) (e.g. [46]). However, the last finding has been challenged [22] as top-down factors (e.g. the task) seem to have a large influence.

Several studies investigated the amount of gaze and mutal gaze in social interactions, which was found to depend on many factors such as personality, sex and distance to the interaction partner (e.g. [3]), the relations and interpersonal attitudes (e.g. [85,39]), and also in various ways on the type and topic of conversation (e.g. [38,37]) resp. task to perform (e.g. [2,57]). For example, people look more while listening than while talking [3] and prefer to look at people who send positive nonverbal signals [26]. Also the gazing differences in adult-adult, adult-child and adult-robot interaction were investigated [74]. Furthermore, there are a lot of intercultural differences (e.g. [121]). Argyle and Cook presented a comprehensive discussion of the complex role of gaze in social interaction [1].

2.3 Facial Expressions

Facial expressions received a great amount of research attention in recent decades, so we consider them in more detail. The *Facial Action Coding System* (FACS) developed by Ekman and Friesen [32] is the most widley used technique to encode and represent facial expressions. A facial expression is decomposed into a set of *Action Units* (AUs) that are directly related to facial muscle movements, thus a representation in terms of AUs describes the visual appearance of a facial expression and does not attribute a specific meaning to it. Nevertheless it can be used as a solid basis for a subsequent interpretation.

In social interactions, the interpretation of a facial expression is more important than its visual appearance, therefore we focus on this interpretation in the following sections. Facial expressions are closely related to emotions. We discuss this in Sect. 2.3.1 and an alternative view that emphasizes the communicative meaning in Sect. 2.3.2.

2.3.1 Emotional Facial Expressions

One basic question is whether emotions and associated facial expressions are universal and innate, or culture-specific and learned. A widely recognized answer is given by Ekman's *neuro-cultural theory of facial expressions of emotion* [27], stating that some emotions are universally tied to particular facial expressions, though there are cultural variations regarding the elicitors of emotions, social display rules for facial expressions, and consequences of emotional arousal. He reported several studies that provide evidence for the universal recognition of the *basic emotions* fear, anger, sadness, disgust, surprise, and happiness across cultures. Ekman [28] also suggested awe, contempt, embarrassment, excitement, guilt, interest, and shame as additional



(a) Stationary scene camera (b) Stationary face camera behind Biron (c) Biron's pan tilt camera

Fig. 1 During the object-teaching user study, videos were recorded from three perspectives. Please refer to Sect. 3.1.

candidates and discussed general charackteristics of basic emotions that might distinguish them from other affective states or non-basic emotions.

Russell [105] questioned this evidence gained from the studies of Ekman *et al.* [34,27,33] and others [61,90,11, 24,82], criticized the experiments on several levels (e.g. forced choice experiments, subject selection, within-subject design, posed expressions, previewing), and suggested several alternative interpretations (e.g. bipolar dimensions, response to a situation, different facial expression categories). Ekman [29] and Izard [62] countered this criticism and defended the conducted studies, but this did not convince Russell of the superiority of Ekman's interpretations over several alternatives [106].

One widely used alternative are *dimensional models* that regard emotions as varying along bipolar, nearly independent dimensions. Mehrabian and Russell [86] presented pleasure, arousal, and dominance as the three fundamental emotional dimensions, largely based on *semantic differential* studies [92,93,112,16]. In later work, Russell [102] reviewed several studies and presented evidence for pleasure and arousal dimensions, whereas the evidence for dominance was not as clear; Russell considered it no longer an affect dimension later on [103] (although further, nonaffective dimensions apparently exist). He also defended this model against methodical criticism (e.g. [49]) and studies suggesting monopolar dimensions (e.g. [114,84]) by criticizing the response format and correcting for a thereby introduced bias [103], though there is some evidence favouring monopolar interpretations, especially showing that positive and negative affect can occur simultaneously (e.g. [122,83,111,58]). Further research led to the *circumplex model of affect* [104, 97], where the dimensions are viewed as systematically interrelated rather than independent. Bradley and Lang [12] developed the *self-assessment manikin* as a simpler alternative to the semantic differential to access affective responses and also propose a pleasure-arousal-dominance model.

Izard [63] recently pointed out that there is still no commonly accepted definition of "emotion", despite broad agreement on several aspects. He suggested that researchers should contextualize their understanding of emotion to clarify its meaning. Widen and Russell [123] added on that by emphasizing the difference between an every day and a scientific concept of emotion.

2.3.2 Communicative Facial Expressions

Fridlund [44] presented the *behavorial ecology view* of faces which is very different from the emotions view discussed above. Facial expressions are regarded as communicative signals that enhance social interaction rather than external displays of internal emotions. No prototypical facial expressions are proposed, as the meaning of a facial display depends heavily on the context. This view is supported by human audience effect studies. Generally, smiles most often occur in social contexts [4,98]. Kraut and Johnston [69] reported that bowlers' smiles were much more related to social interaction with the people around than to scoring a strike or spare. Similar results were obtained for people experiencing good or bad weather and, too a weaker degreee, for fans watching a hockey game. Fridlund *et al.* [45] reported that people smiled more when imagining high-sociality situations compared to low-sociality ones and that the degree of smiling was little related to their happiness. Fridlund [43] also showed that people watching an amusing video smiled more when a friend was present, or even when they were told that a friend was in a room nearby, each compared to watching alone. Very similar results were obtained by Chovil [19] for the facial display of subjects hearing about close call events. Bavelas *et al.* [8] presented evidence that the motor mimicry of subjects observing apparently painfully injured victims can reasonly be interpreted as communicative act. Brightman *et al.* [13,14] found that observing judges could easily tell whether videotaped subjects were eating sweet or salty sandwiches when the subjects were in company, but not when they were alone.

However, there is also evidence that social context can inhibit the display of negative facial expressions (e.g. [66,

68,64]). Jakobs *et al.* [64] interpreted their results for sad faces as being largely compatible with the idea of display rules as suggested by Ekman [27] and less supportive for the behavorial ecology view [44], but also admitted that the experimental setting might have influenced the subjects against behaving as expected by this view. Ekman *et al.* [31,28] emphasized that facial expressions also occur when people are alone and not imagining others, which questions a sole communicative role. Furthermore, Ekman [30] defended the emotional view of facial expressions and argued that they were not deliberately made to communicate, although emotions play a role in interaction. Parkinson [95] compared Ekman's [27] and Fridlund's [44] approaches and reviewed both in terms of theory, evidence and consequences. He concluded that neither approach can account for all the available evidence: many results cannot be explained by a pure emotions view, the behavorial ecology view covers a wider range of phenomena, but is too imprecise regarding the exact relation of the facial display to social motives and audience effects and cannot explain all emotional displays; further research should aim at a comprehensive theory of facial movements and state these relations more precisely.

2.4 Conclusion

The display and meaning of head gestures, eye gaze, and facial expressions depend heavily on the interaction context and is very complex and multifaceted, especially when one considers the interrelations between these signals (e.g. [108,9,20,56]). We suggest that these three kinds of FCSs should not be treated as individual modules, but should be considered in combination altogether. Due to the complexity we doubt that a comprehensive, general purpose interpretation of FCSs by robots interacting with humans will be feasible in the near future. Therefore, we think that a pragmatic simplification and focusing on different, specific interaction scenarios will remain necessary and beneficial for the midterm development of this aspect of human-robot interaction. (Thus, the overall interpretation capabilities of a robot rather might arise from the combination of several subsystems that are dedicated to specific tasks and contexts than from one general purpose system for FCS interpretation.) Sect. 3 presents our approach to one such scenario. To discover which FCSs actually occur in a specific interaction scenario, we prefer a data-driven approach to an a priori modelling, as it might be very to difficult to anticipate which FCSs are the most prominent ones in a certain context. An important issue for a classification of FCSs is the acquisition of reliable ground truth data. We suggest a definition of ground truth in terms of the objectively ascertainable interaction situation instead of the visual appearance of the face, because this circumvents some typical problems (Please see Sect. 3.2). Furthermore, we suggest an FCS interpretation in broader clusters instead of finegrained categories, because we expect the former to generalize better to other interaction scenarios due to the context-dependence of FCSs.

3 Valence Recognition

In order to find relevant FCSs that actually occur in typical human-robot interactions, we evaluated videos of two user studies, where several subjects showed around a robot in an apartment [75] resp. taught the names of several objects to a robot [76]. (Neither study was related to FCSs originally.) Not surprisingly, it turned out that typical facial expressions of basic emotions [27] rarely occured, which is in accordance with the experiences of Caridakis *et al.* [17]. In many cases, several human raters found the observed facial displays not very clearly visible but rather subtle and difficult to interpret in terms of exact categories. (More than 50 categories were named by the observers.) The agreement of the raters about the best suited category for a particular facial display was often poor, also an expedient and comprehensive set of those categories was not defined.

Nevertheless, the object-teaching scenario was found to be well-suited for FCS studies in general because of the frequent occurrence of FCSs, despite their difficult interpretation in terms of precise categories. However, a classification into broader clusters (e.g. positive vs. negative) achieved much higher agreement among the human raters. This motivates the object-teaching scenario described in Sect. 3.1 and the valence recognition approach presented in Sect. 3.2.

3.1 Object-Teaching Scenario

We conducted a user study with 11 subjects (five female and six male) in an object-teaching scenario. The subjects were instructed to teach the names of several manipulable objects to the robot "Biron"[1][53] and to verify the learning effect. This was a Wizard of Oz study where we controlled the robot's behavior (what to say when, where to look, when to recognize the object correctly, when to misunderstand the subject, etc.). Of course the subjects did not know this, but assumed autonomous operation of the robot whose object recognition capabilities were to be evaluated, whereas in fact the study was about provoking authentic, spontaneous FCSs for later analysis. The subjects were not aware that FCSs were of any interest during the study, which we regard as a necessary prerequisite when authentic FCSs are investigated. A pre-study confirmed that the subjects were likely to display FCSs spontaneously nevertheless. It was not specified how the objects should be taught (e.g. pointing to them, taking them in hand, etc.), but the subjects were asked to

[1] Bielefeld Robot Companion

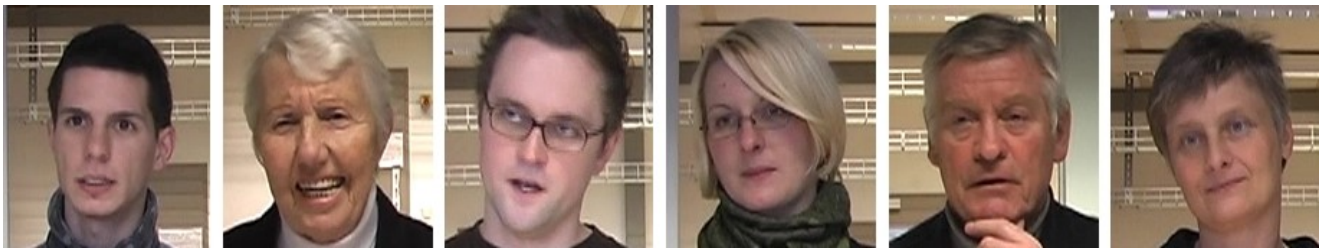

Fig. 2 Example snapshots of the facial communicative signals (FCSs) that occured in the object-teaching user study.

interact with the robot at their convenience, in order to perform the given task. They were aware that the robot understood speech and could see them. The robot interacted with the subjects by voice production and movements of its pan tilt camera. Per subject, two counterbalanced runs were performed: a "good" one where most objects were classified correctly, and a "bad" one with very frequent misclassifications. Each run was videotaped from three different perspectives, as depicted in Fig. 1. Some examples of the displayed FCSs are shown in Fig. 2. An evaluation of the videos showed that the interactions were highly structured and can be subdivided into four phases:

1. *present:* the subject presented the object to Biron and said its name or asked for the name
2. *waiting:* the subject waited for the answer of the robot (not mandatory)
3. *answer:* the robot answered the subject (e.g. classifying the object or asking a question)
4. *react:* the subject reacted to the answer of the robot

The videos of the stationary face camera (Fig. 1b) were manually annotated according to a predefined coding scheme to mark the beginning and end of each phase for each object-teaching scene, also the voices of the subject and Biron were transcribed. These annotations were used in the evaluation of the human valence recognition performance, as reported in Sec. 3.2. Details on this object-teaching study and the recorded video corpus can be found in [70].

3.2 Human Valence Recognition Performance

In order to perform a feedback interpretation study, we selected those object-teaching scenes where Biron answered to its human interaction partner with a concrete object name, which could either be the correct name (*success* scene) or a wrong one (*failure* scene). In total, there are 221 *success* scenes and 226 *failure* scenes in the video database. The task of the subjects in this study was to decide whether or not Biron classified an object correctly, only by looking at the face of its interaction partner during the relevant part of the interaction. Thus, they should interpret the FCSs given by those people in terms of *valence*. Before we report the details and results of this study, we summarize our motivation for feedback interpretation in terms of valence in an object-teaching scenario:

- Reliable ground truth data is available: one knows for sure whether the answer of the robot is correct or not.
- The subjects showed spontaneous FCSs (as opposed to posed ones that were used in many studies, e.g. [27]).
- Although not guaranteed, the subjects are likely to show some prominent FCSs as reaction to the robot's answer, because they want the robot to classify the objects correctly. This was confirmed by our preliminary studies and also by the studies of Barkhuysen *et al.* [6].
- The roughly predictable interaction structure (present, waiting, answer, and react phases) can simplify the design of the robot's behavior in autonomous interactions.
- It is a realistic scenario and a challenging problem for the automatic recognition of FCSs.

It is important to note that this definition of ground truth is significantly different from the most widely used definition. Usually, the ground truth label is directly connected to the visual appearance of the face (e.g. displaying a "happy" face) and is gainend by human judges interpreting the face, self-report studies, or via posing FCSs on request. In our case, the ground truth label is defined by the objectively ascertainable *interaction situation*, namely whether the robot recognized the object correctly or not, regardless of the facial display in this situation.[2] Thus the research question addressed here is not the recognition of FCSs in themselves, but their interpretation as *feedback* in a concrete interaction situation. This is similar to the approach of Barkhuysen *et al.* [6], who performed a problem detection study where subjects judged videos showing people interacting with an oral train timetable dialog system.

3.2.1 Feedback Interpretation Study

We randomly selected four *success* and four *failure* scenes of each of the 11 interaction partners shown in the object-

[2] Thus, this definition of ground truth circumvents typical problems of the common approaches mentioned above: There is no dependence on the necessarily subjective impressions of human raters, subjects do not need to remember the intended meaning of their various FCS displays during an interview after the experiment, and the displayed FCSs are spontaneous and not posed on request.



Fig. 3 Example snapshots of the videos used in the feedback interpretation study. Please refer to Sect. 3.2.1.

teaching videos (Sect. 3.1), summing to 88 scenes in total. These scenes were judged by 44 subjects (15 female and 29 male) whose task was to decide whether Biron classified the objects correctly. Only the task-relevant parts of the videos were shown, i.e. a video starts when Biron starts to utter the object name and ends when the annotated reaction phase of the subject is finished (Sect. 3.1); the sound was removed from all videos. The amount of displayed context information was varied in two respects: showing the full scene vs. showing the face only (varying visual context), each combined with playing the video over the full length vs. playing only to first half of it (varying temporal context). Fig. 3 shows several example snapshots of these videos. The subjects were allowed to watch a scene several times.

3.2.2 Results

The results for the four different context conditions are summarized in Tab. 1 and Tab. 2:

- *all*: average over all four context conditions
- *fs-fl*: showing the full scene and full video length
- *fs-hl*: showing the full scene and first half of video
- *of-fl*: showing only the face and full video length
- *of-hl*: showing only the face and first half of video

The visual context helped significantly in the classification of half-length videos (t-test, $p < 0.01$), but only very slightly and not significantly in the classification of full-length videos ($p < 0.61$). The temporal context was found to be more important as it significantly improved the classification in both cases full-scene ($p < 0.03$) and face-only ($p < 0.001$) videos. On average, *failure* videos were easier correctly classified than *success* videos ($p < 0.02$). The variance over observing subjects was high, but the variance over the judged videos even higher. This is due to the large variability in the expressiveness regarding FCSs of the teaching subjects shown in the videos. For instance, the person in the best recognized video shows a clear nodding, whereas the subject in the most poorly recognized video shows hardly any FCS at all. For further details on the experimental procedure and results and a comparison to the results of a similar study by Barkhuysen *et al.* [6], please refer to [70].

3.3 Conclusion

The evaluation of the human valence recognition performance in an object-teaching scenario showed that, given a sufficient length of the video sequence, humans can interpret FCSs in terms of valence almost equally well when only the face is shown, compared to the display of the full scene. Thus the visual context is not that important, the restriction to the face is in principle sufficient, which allows automatic recognition approaches to be reasonably constrained to face processing in this scenario.

However, the average human classification performance in this case (only face, full video length) was only 82% and showed a high variance depending on the observed video. This is comparatively low for a two-class classification problem and illustrates its difficulty. Due to the large variations in the expressiveness of different people, a high variance is expected for automatic recognition approaches as well.

4 Automatic Recognition of Facial Communicative Signals

We briefly discuss related works on the automatic recognition of FCSs in Sect. 4.1, before we describe the utilized features in Sect. 4.2 and introduce a static and a dynamic recognition approach to this problem in Sect. 4.3 and Sect. 4.4, respectively. Both approaches were evaluated on all *success* and *failure* videos of the database introduced in Sect. 3.1.

4.1 Related Work

Murphy-Chutorian and Trivedi [88] presented a comprehensive survey on visual head pose estimation. They classified the existing approaches into eight categories, reviewed their theory and compared the achieved results. We exemplarily list some state of the art approaches below. Wang and Sung [120] used geometric relations of eyes and mouth corners to estimate the 3D pose of a head, whereby they utilized an EM-algorithm to adapt to different persons and facial expressions. Zhao *et al.* [128] introduced a 3D head tracking method that uses image registration based on SIFT fea-

sub-	all videos		success videos		failure videos	
set	mean	std	mean	std	mean	std
all	79.1	8.2	75.8	11.9	82.4	12.0
fs-fl	83.4	12.8	80.2	16.8	86.6	16.2
fs-hl	78.2	8.1	75.0	12.6	81.4	15.3
of-fl	82.0	11.1	78.1	16.3	86.0	13.5
of-hl	72.8	9.9	69.8	15.9	75.8	15.9

Table 1 Mean value and standard deviation of the classification accuracy for different context conditions (distribution over observing subjects). Please refer to Sect. 3.2.2.

sub-	all videos		success videos		failure videos	
set	mean	std	mean	std	mean	std
all	79.1	17.9	75.8	19.4	82.4	15.8
fs-fl	83.4	18.1	80.2	21.1	86.6	14.1
fs-hl	78.2	24.0	75.0	27.2	81.4	20.1
of-fl	82.0	19.1	78.1	21.2	86.0	16.1
of-hl	72.8	23.9	69.8	25.8	75.8	21.7

Table 2 Mean value and standard deviation of the classification accuracy for different context conditions (distribution over judged videos). Please refer to Sect. 3.2.2.

tures [77]. Baker *et al.* [5] described approaches for non-rigid head tracking using 2D and 3D active appearance models (AAMs) [23]. Ma *et al.* [79] presented a non-linear regression method that uses a relevance vector machine [116] to learn the relations between the position of facial feature points and the 2D head pose.

Morimoto and Mimica [87] reviewed several eye tracking approaches and concluded that recent nonintrusive state of the art methods can yield sufficient accuracy for many applications. Yoo *et al.* [126], for instance, reported a real-time eye tracking system for people sitting in front of a computer monitor. Like many other approaches, they used infrared light, which might be a disadvantage in some use cases. Wang and Sung [119] presented a system that uses the geometric relations of iris and eye corners, evaluated in a zoomed-in image of one eye, to robustly estimate the eye gaze. A stereo vision system for real-time head pose and eye gaze estimation by means of 3D eye corners and pupils tracking via template matching was described by Newman *et al.* [89]. Ishikawa *et al.* [59] used an AAM to locate the eye region and a subsequent ellipse fitting and template matching for gaze estimation, whereas Ivan [60] directly utilized AAMs to model the eye. Varchmin *et al.* [118] developed a system that combines eigeneye analysis, nose and mouth detection (for head pose estimation), and a series of neural networks to estimate the gaze direction of a user.

Fasel and Luettin [40] and also Pantic and Rothkrantz [94] presented surveys on facial expression recognition approaches. Many researchers investigated a classification into discrete categories, most often basic emotions [27]. Buenaposada *et al.* [15] built linear subspace deformation and illumination models and used a nearest-neighbor-based classifier for that purpose, whereas Lanitis *et al.* [73] used flexible models of shape and gray-level, also the related AAMs were utilized in other works [25,100], as well as many other techniques, e.g. haar-like features and dynamic binary patterns [124] and local facial feature deformations [110]. Some works considered other or additional categories (e.g. [109, 65,125]), including complexer mental classes [35].

Also the recognition of action units (AUs) [32] has been investigated, for instance by Bartlett *et al.* [7], who classified 20 AUs by means of gabor filters and support vector machines (SVMs), Tian *et al.* [115], who presented a system that can recognize 16 AUs via geometric facial feature modeling and two neural network classifiers for the upper and lower part of the face, and several others (e.g. [117,78, 21]). A few works consider the recognition of facial expressions in terms of emotional dimensions. Caridakis *et al.* [17] and Fragopanagos and Taylor [42] investigated the recognition of valence and activation level with neural networks. Gunes and Pantic [52] used hidden markov models (HMMs) and SVMs for the continuous prediction of five dimensions (arousal, expectation, intensity, power and valence). In total, however, the recognition of valence and other dimensions is not as intensively researched as the classification of discrete emotion categories.

Most early works considered posed facial expressions. Nowadays, there is a growing interest in authentic, spontaneous facial expressions, that are quite different from posed ones. Valstar *et al.* [117] showed that genuine and posed smiles can be distinguished automatically. Sebe *et al.* [110] investigated the classification of authentic basic emotions in a video kiosk scenario. Bartlett *et al.* [7] performed AU recognition on a database of subjects engaged in social or political discussions. Zeng *et al.* [127] presented a comprehensive survey on this topic.

4.2 Feature Extraction

We used active appearance models (AAMs) [23,80] for feature extraction, because it is a state of the art approach for facial feature detection and tracking and has been successfully utilized for the recognition of head poses and gestures, eye gaze, and facial expressions (Sect. 4.1), making it well suited for the recognition of FCSs in our scenario. An AAM requires a set of training images with annotated feature points. From this training data, linear models of shape and texture are constructed via principal component analysis (PCA). The shape model captures the variance of the feature point positions in the training data (aligned via a procrustes analysis), whereas the texture model represents the shape-aligned gray-level variance of the training images. Due to the feature point spreading over the inner face area, the AAM can capture facial expressions very well. The model includes a global scaling and rotational alignment. Together

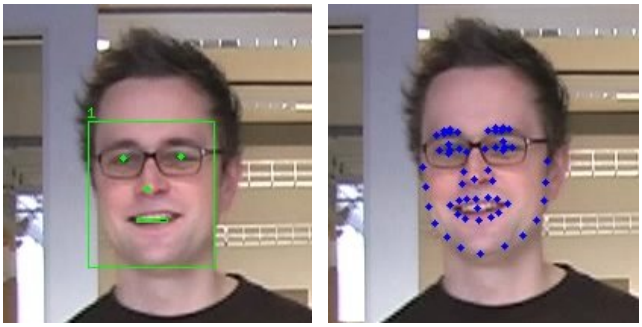

Fig. 4 Examples for the utilized face detection and feature extraction methods. Please refer to Sect. 4.2

with feature point displacements over time caused by head movements, this allows the AAM to also capture head gestures, provided that out-of-plane rotations are limited such that all feature points are still visible. This is usally the case in the investigated scenario, as the subjects mainly direct their attention towards the robot. Although not explicitly modeled, the AAM can also account for eye gaze to some degree: pupil position and gaze direction are roughly captured by the position of and texture under the five feature points that are dedicated to each eye. Moreover, changes in gaze direction are accompanied by head movements most of the time.

The fitting to a new image is achieved by means of an iterative search algorithm which requires a suitable initialization. Therefore, we applied the face detection technique developed by Castrillón *et al.* [18], that detects the face bounding box and the positions of eyes, mouth, and nose. This information is used to initialize the AAM fitting, based on the method described by Rabie *et al.* [99]. Fig. 4 shows an example of an AAM fitted to an input image from our database. The appearance parameter vector (representing both shape and texture parameters) of the AAM is used as feature vector in the classification of FCS, as reported in the following sections. As person-specific AAMs usually yield a considerably better fitting performance than generic ones [50], we built an individual AAM for each of the 11 persons in the database, each trained with approximately 200 annotated images. For the basic investigations presented in this paper, person-specific AAMs are acceptable. However, as generic models are important for human-robot interactions with naive users outside laboratories, we will investigate the usage of generic AAMs in future work.

4.3 A Static Recognition Approach

We first evaluated the classification performance of a static SVM classifier that operates on single frame level and does not consider temporal dynamics. The evaluation was conducted on the video sequences showing the task-relevant parts of the *success* and *failure* scenes of all subjects, according to the annotations described in Sect. 3.2. We performed a person-specific classification using leave-one-out cross-validation over the videos of each subject, where all frames of all but one video were used as training data, while the frames of the remaining video were used as test data. The AAM parameter vectors (Sect. 4.2) of all frames were independently classified and fused via majority voting to generate the classification result for a video.

Tab. 3 summarizes the results for different selections of parameters (rbf parameter σ and regularization cost C of the SVM's RBF kernel). The first row ("mv-cross-val") shows the classification rates for the best parameters found via a ten-fold cross-validation on the respective training data prior to the classification of each test video. We were also interested in the stability of these parameters and used their mean values (over all videos of a subject) in all classifications of a subject. The second row ("mv-average") shows that this yielded comparable, even slightly improved classification rates, thus stable parameters can be chosen for each person.

In spite of the approach being simple already, we further simplified the features by using only the mean AAM parameter vector of each video as representation of the whole video. Surprisingly, this improved the classification performance slightly, as listed in the third ("mean-cross-val") and fourth ("mean-average") row of Tab. 3, which summarizes the results for parameters found via a ten-fold cross-validation and the mean parameters, respectively. An investigation of this finding revealed that often only a subsequence of a video is actually discriminative in terms of *success* and *failure*, although the videos were presegmented to contain only relevant information. Hence, a high number of potentially irrelevant frames can disturb the majority voting such that it is not superior to a simple averaging over frames which still can capture some essential information.

For further details on the experiments and more results, also for other features (gabor energy filters and raw images), please refer to [71]. Though the task is difficult, the recognition rates achieved by this simple approach are rather low for a two-class problem, but can serve as a first baseline for automatic recognition methods nevertheless. Because of the principle limitations of static methods due to neglecting temporal information, we did not delve into this further, but considered a more sophisticated dynamic approach instead, which is introduced in the next section.

4.4 A Dynamic Recognition Approach

Motivated by the observation that often only a subpart of a video sequence appears to be relevant for the classification and the general assumption that the temporal dynamics are important for the interpretation of spontaneous fa-

variants	all scenes mean	all scenes std	success mean	success std	failure mean	failure std
mv-cross-val	74.2	11.0	63.0	19.1	82.5	14.3
mv-average	75.4	10.1	64.1	19.3	84.0	12.5
mean-cross-val	76.0	11.5	70.3	19.2	79.2	13.3
mean-average	77.6	11.6	73.5	16.7	80.3	10.1

Table 3 Mean classification accuracy and standard deviation for the static SVM classifier for all videos, only *success* and only *failure* videos. For an explanation of the listed variants, please refer to Sect. 4.3.

feature variants	all scenes mean	all scenes std	success mean	success std	failure mean	failure std
cross-val	79.5	10.2	79.8	13.1	77.2	17.0
median-scenes	86.5	7.7	85.8	14.1	85.7	8.6
median-persons	74.3	10.7	73.3	19.1	77.8	23.6

Table 4 Mean classification accuracy and standard deviation for the dynamic classifier for all videos, only *success* and only *failure* videos. For an explanation of the listed variants, please refer to Sect. 4.4.

cial behavior (e.g. [29]), we investigated the recognition of FCSs by means of matching subsequences of a test video to certain reference subsequences. An exhaustive search on the training videos is performed to find subsequences (of a given minimal and maximal length) in these videos that are characteristic for either *success* or *failure*. This is done by a ranking of all considered subsequences according to a score value that favors subsequences which are very similar to other subsequences of the same class, but rather different from even the most similar subsequences of the other class. This scoring function is related to the Fisher criterion [41], which minimizes the within scatter while maximizing the between scatter of data from two classes to find an optimal discriminant function. The idea is that subsequences with high score values constitute characteristic prototypes of their class while being dissimilar to any subsequence of the other class. The similarity of two subsequences is computed by means of dynamic time warping [107] over the AAM parameter vector sequences, where the euclidean distance between two AAM vectors (each associated with a particular frame) is used in the computation. A certain number of the best ranked subsequences is chosen as reference subsequences for each class.

During the classification, the reference subsequences are matched to an input video. More precisely, all subsequences (of a given minimal and maximal length) of the input video are compared to the reference subsequences in a k-nearest-neighbor-based classification approach, where dynamic time warping is used as distance measure again.[3] Based on the distances to the nearest reference subsequences, a classification score is computed for each class. The input video is classified into the class with the highest classification score. Our visual impression is that while some people display positive and negative valence with approximately equal expressiveness, others show a clear bias, meaning that the absence of failure signs can reasonably be interpreted as success, or vice versa. In terms of the AAM features, this results in a different range of variation for subsequences of the two classes. This motivates the introduction of an adjustment factor on the classification scores of one class, serving as an a priori bias, which facilitates a fair comparability of these score values. This adjustment factor is optimized on the training data via cross-validation and was found to improve the classification rates considerably.

Tab. 4 shows the results for a person-specific classification via leave-one-out cross-validation over the videos (row "cross-val"), likewise to the static SVM classification, compared to which the average classification rate improved and is now close to the human performance. The described dynamic classifier involves several parameters (e.g. the number of reference subsequences and distance values to consider in the score calculations), which are optimized via cross-validation on the training data. When the median values of these parameters were used in all classifications of a subject, the classification rate improved notably (row "median-scenes"), whereas it dropped when a second median operation was performed to use the same parameters for all subjects (row "median-persons"). This indicates that stable parameters can be chosen for each person, but also that these parameters are person-specific and do not generalize well over subjects.

Visual inspection of the computed reference subsequences showed that all three kinds of FCSs that were considered in this paper (head gestures, eye gaze, and facial expressions) occured, with in part large variations between subjects. In many cases, the subjects tend to use facial expressions and verbal correction to signal *failure*, but head gestures and eye gaze (primarily a gaze direction shift from the robot to the object table) to indicate *success*. Especially the last FCS appears to be task-dependent. Fig. 5 depicts some typical example images taken from the most discriminative reference subsequences of four subjects. The full details of this dynamic recognition technique, further results, and a comparison to related methods (e.g. [91]) are presented in a paper that is currently under review. This basic approach was found to be promising and shall be further investigated and improved in future work.

[3] The start and end points of the test sequence are given by the manual annotation of the database (Sect. 3.1), as this paper focuses on the principal investigation of FCSs in the described scenario. Nevertheless, an automatic determination of these segment borders as needed for the online-classification on the robot is also possible, using the robot's system state and simple heuristics. Approximate segment borders are sufficient, as a search for the best-matching subsequences is performed anyway. Also an incremental evaluation without a fixed end point is possible efficiently. However, the details of such a robot system are beyond the scope of this paper and will be presented elsewhere.



Fig. 5 Example images from the selected reference subsequences. Top row: signaling *success* via head gestures (left) and gaze direction (right). Bottom row: signaling *failure* via facial expressions. Please refer to Sect. 4.4.

5 Conclusion

We discussed head gestures, eye gaze, and facial expressions in their function as facial communicative signals (FCSs) and suggested that due to their complex and partially disputed meaning, the research on automatic recognition of FCSs in human-robot interaction should continue to focus on specific interaction scenarios in the midterm, as a comprehensive, general purpose interpretation of FCSs seems not realizable in the near future. We described an interaction task where subjects taught objects to a robot as one example of such a scenario and evaluated the human performance in the recognition of FCSs as nonverbal feeback in terms of valence. This turned out to be a difficult task, as the average human recognition accuracy was only about 82%. The ground truth data was implicitly defined by the interaction situation, in contrast to its usual definition in terms of the face's visual appearance (often judged by human raters).

Furthermore, we reported a simple static approach for the automatic recognition of FCSs in this scenario using active appearance models (AAMs) for feature extraction and a support vector machine (SVM) for classification, whose results served as a first baseline. Subsequently, we introduced a more sophisticated approach that considers the temporal dynamics in the face videos via dynamic time warping and performs a classification based on reference subsequence selection. This method outperformed the static baseline approach and yielded classification rates comparable to the human performance in a person-dependent classification.[4]

The presented investigations are preliminary steps towards the fully automatic recognition of FCSs in task-oriented human-robot interaction. Future work will focus on the generalization to new subjects, which is very challenging due to the large variations in the display of FCSs by different persons, especially given the comparatively small number of subjects in the current database. The goal is to provide robots with the ability to interpret the nonverbal feedback given by facial communicative signals of their interaction partners in certain situations.

[4] However, it is not clear to which degree the humans performed a person-dependent or -independent classification. At least some adaptation to the shown people took place in the course of the experiment.

Acknowledgements Christian Lang gratefully acknowledges the financial support from Honda Research Institute Europe for the project "Facial Expressions in Communication".

References

1. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press (1976)
2. Argyle, M., Graham, J.A.: The central europe experiment: Looking at persons and looking at objects. J Nonverbal Behav **1**(1), 6–16 (1976)
3. Argyle, M., Ingham, R.: Gaze, mutual gaze, and proximity. Semiotica **6**(1), 32–49 (1972)
4. Bainum, C.K., Lounsbury, K.R., Pollio, H.R.: The development of laughing and smiling in nursery school children. Child Dev **55**(5), 1946–1957 (1984)
5. Baker, S., Matthews, I., Xiao J. Gross, R., Kanade, T., Ishikawa, T.: Real-time non-rigid driver head tracking for driver mental state estimation. In: 11th World Congress Intell Transp Syst (2004)
6. Barkhuysen, P., Krahmer, E., Swerts, M.: Problem detection in human-machine interactions based on facial expressions of users. Speech Commun **45**(3), 343–359 (2005)
7. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. In: International Conference on Automatic Face and Gesture Recognition, pp. 223–230 (2006)
8. Bavelas, J.B., Black, A., Lemery, C.R., Mullett, J.: I show how you feel: Motor mimicry as a communicative act. J Pers Soc Psychol **50**(2), 322–329 (1986)
9. Bindemann, M., Burton, A.M., Langton, S.R.H.: How do eye gaze and facial expression interact? Vis Cognition **16**(6), 708–733 (2007)

10. Birdwhistell, R.: Kinesics and Context: Essays on Body Motion Communication. University of Pennsylvania Press (1970)
11. Boucher, J.D., Carlson, G.E.: Recognition of facial expression in three cultures. J Cross-Cultural Psychol **11**, 263–280 (1980)
12. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. J Behav Therapy Exp Psychiatry **25**(1), 49–59 (1994)
13. Brightman, V.J., Segal, A.L., Werther, P., Steiner, J.: Ethologic study of facial expressions in response to taste stimuli. J Dental Res **54**, L141 (Abstract) (1975)
14. Brightman, V.J., Segal, A.L., Werther, P., Steiner, J.: Facial expression and hedonic response to face stimuli. J Dental Res **56**, B161 (Abstract) (1977)
15. Buenaposada, J.M., Muñoz, E., Baumela, L.: Recognising facial expressions in video sequences. Pattern Analysis & Applications **11**(1), 101–116 (2008)
16. Bush, L.E.: Individual differences multidimensional scaling of adjectives denoting feelings. J Pers Soc Psychol **25**(1), 50–57 (1973)
17. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: International Conference on Multimodal Interfaces, pp. 146–154 (2006)
18. Castrillón, M., Déniz, O., Hernández, M.: The encara system for face detection and normalization. Lecture Notes in Computer Science **2652**, 176–183 (2003)
19. Chovil, N.: Social determinants of facial displays. J of Nonverbal Behav **15**(3), 141–154 (1991)
20. Cline, M.G.: The perception of where a person is looking. American J Psychol **80**(1), 41–50 (1967)
21. Cohn, J., Reed, L., Ambadar, Z., Xiao, J., Moriyama, T.: Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In: International Conference on Systems, Man and Cybernetics, pp. 610–616 (2004)
22. Cooper, R.M.: The effects of eye gaze and emotional facial expression on the allocation of visual attention. Ph.D. thesis, University of Stirling, Department of Psychology (2006)
23. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. Pattern Analysis Machine Intelligence **23**(6), 681–685 (2001)
24. Ducci, L., Arcuri, L., W/ Georgis, T., Sineshaw, T.: Emotion recognition in ethiopia: The effect of familiarity with western culture on accuracy of recognition. J Cross-Cultural Psychol **13**, 340–351 (1982)
25. Edwards, G., Cootes, T., Taylor, C.: Face recognition using active appearance models. In: H. Burkhardt, B. Neumann (eds.) European Conference on Computer Vision, vol. 2, pp. 581–695. Springer (1998)
26. Efran, J.S.: Looking for approval: Effects on visual behavior of approbation from persons differing in importance. J Pers Soc Psychol **10**(1), 21–25 (1968)
27. Ekman, P.: Universals and cultural differences in facial expressions of emotion. Nebraska Symp Motivation **19**, 207–283 (1971)
28. Ekman, P.: An argument for basic emotions. Cognition Emot **6**(3 & 4), 169–200 (1992)
29. Ekman, P.: Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. Psychol Bull **115**(2), 268–287 (1994)
30. Ekman, P.: Should we call it expression or communication? Innovation **10**(4), 333–344 (1997)
31. Ekman, P., Davidson, R.J., Friesen, W.V.: The duchenne smile: Emotional expression and brain physiology ii. J Pers Soc Psychol **58**(2), 342–353 (1990)
32. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
33. Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., Scherer, K., Tomita, M., Tzavaras, A.: Universals and cultural differences in the judgments of facial expressions of emotion. J Pers Soc Psychol **53**(4), 712–717 (1987)
34. Ekman, P., Sorenson, E.R., Friesen, W.V.: Pan-cultural elements in facial displays of emotion. Science **164**, 86–88 (1969)
35. El Kaliouby, R., Robinson, P.: Real-time inference of complex mental states from facial expressions and head gestures. Transactions Robot **23**(5), 991–1000 (2007)
36. Emery, N.J.: The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev **24**(6), 581–604 (2000)
37. Exline, R., Gray, D., Schuette, D.: Visual behavior in a dyad as affected by interview content and sex of respondent. J Pers Soc Psychol **1**(3), 201–209 (1965)
38. Exline, R.V.: Explorations in the process of person perception: visual interaction in relation to competition, sex, and need for affiliation. J Pers **31**(1), 1–20 (1963)
39. Exline, R.V., Winter, L.C.: Affect, cognition and personality, chap. Affective relations and mutual glances in dyads. Springer, New York (1965)
40. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognit **36**, 259–275 (2003)
41. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann Eugen **7**, 179–188 (1936)
42. Fragopanagos, N., Taylor, J.: Emotion recognition in human-computer interaction. Neural Netw **18**(4), 389–405 (2005)
43. Fridlund, A.J.: Sociality of solitary smiling: Potentiation by an implicit audience. J Pers Soc Psychol **60**(2), 229–240 (1991)
44. Fridlund, A.J.: Human facial expression: An evolutionary view. Academic Press, San Diego, CA (1994)
45. Fridlund, A.J., Sabini, J.P., Hedlund, L.E., Schaut, J.A., Shenker, J.I., Knauer, M.J.: Audience effects on solitary faces during imagery: Displaying to the people in your head. J Nonverbal Behav **14**(2), 113–137 (1990)
46. Friesen, C.K., Moore, C., Kingstone, A.: Does gaze direction really trigger a reflexive shift of spatial attention? Brain Cognition **57**(1), 66–69 (2005)
47. Goodwin, M.H., Goodwin, C.: Gesture and coparticipation in the activity of searching for a word. Semiotica **62**(1–2), 51–76 (1986)
48. Graf, H.P., Cosatto, E., Strom, V., Huang, F.J.: Visual prosody: Facial movements accompanying speech. In: International Conference on Automatic Face and Gesture Recognition, pp. 396–401 (2002)
49. Green, R.F., Goldfried, M.R.: On the bipolarity of semantic space. Psychol Monogr Gen Appl **79**(6, Whole No. 599), 31 (1965)
50. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image Vis Comput **23**(12), 1080–1093 (2005)
51. von Grünau, M., Anston, C.: The detection of gaze direction: A stare-in-the-crowd effect. Percept **24**(11), 1297–1313 (1995)
52. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: International Conference on Intelligent Virtual Agents, pp. 371–377 (2010)
53. Haasch, A., Hohenner, S., Hüwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G.A., Fritsch, J., Wrede, B., Sagerer, G.: Biron - the bielefeld robot companion. In: E. Prassler, G. Lawitzky, P. Fiorini, M. Haegele (eds.) International Workshop on Advances in Service Robotics, pp. 27–32 (2004)
54. Hadar, U., Steiner, T.J., Grant, E.C., Rose, F.C.: Kinematics of head movements accompanying speech during conversation. Hum Mov Science **2**(1–2), 35–46 (1983)

55. Heylen, D.: Challenges ahead: Head movements and other social acts in conversations. In: Joint Symposium on Virtual Social Agents, pp. 45–52 (2005)
56. Heylen, D.: Head gestures, gaze and the principles of conversational structure. Int J Humanoid Robot **3**(3), 1–27 (2006)
57. Hugot, V.: Eye gaze analysis in human-human interactions. Master's thesis, KTH Royal Institute of Technology, School of Computer Science and Communication, Stockholm, Sweden (2007)
58. Huppert, F.A., Whittington, J.E.: Evidence for the independence of positive and negative well-being: Implications for quality of life assessment. Br J Health Psychol **8**, 107–122 (2003)
59. Ishikawa, T., Baker, S., Matthews, I., Kanade, T.: Passive driver gaze tracking with active appearance models. In: 11th World Congress on Intelligent Transportation Systems (2004)
60. Ivan, P.: Active appearance models for gaze estimation. Master's thesis, Vrije Universiteit Amsterdam, Faculty of Sciences, Business Mathematics & Informatics (2007)
61. Izard, C.E.: The Face of Emotion. Appleton-Century-Crofts (1971)
62. Izard, C.E.: Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychol Bull **115**(2), 288–299 (1994)
63. Izard, C.E.: The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. Emot Rev **2**(4), 363–370 (2010)
64. Jakobs, E., Manstead, A.S.R., Fischer, A.H.: Social context effects on facial activity in a negative emotional setting. Emot **1**(1), 51–69 (2001)
65. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: ACM International Conference on Multimedia, pp. 677–682 (2005)
66. Kleck, R.E., Vaughan, R.C., Cartwright-Smith, J., Vaughan, K.B., Colby, C.Z., Lanzetta, J.T.: Effects of being observed on expressive, subjective, and physiological responses to painful stimuli. J Pers Soc Psychol **34**(6), 1211–1218 (1976)
67. Kobayashi, H., Kohshima, S.: Unique morphology of the human eye. Nat **387**, 767–768 (1997)
68. Kraut, R.E.: Social presence, facial feedback, and emotion. J Pers Soc Psychol **42**(5), 853–863 (1982)
69. Kraut, R.E., Johnston, R.E.: Social and emotional messages of smiling: An ethological approach. J Pers Soc Psychol **37**(9), 1539–1553 (1979)
70. Lang, C., Hanheide, M., Lohse, M., Wersing, H., Sagerer, G.: Feedback interpretation based on facial expressions in human-robot interaction. In: International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 189–194 (2009)
71. Lang, C., Wachsmuth, S., Wersing, H., Hanheide, M.: Facial expressions as feedback cue in human-robot interaction - a comparison between human and automatic recognition performances. In: Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), pp. 79–85 (2010)
72. Langton, S.R.H., Watt, R.J., Bruce, V.: Do the eyes have it? cues to the direction of social attention. Trends Cogn Sciences **4**(2), 50–59 (2000)
73. Lanitis, A., Taylor, C.J., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. Pattern Analysis Machine Intelligence **19**(7), 743–756 (1997)
74. Lohan, K.S., Vollmer, A.L., Fritsch, J., Rohlfing, K., Wrede, B.: Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations? In: International Workshop on Social Signal Processing (2009)
75. Lohse, M., Hanheide, M.: Evaluating a social home tour robot applying heuristics. In: Workshop Robots as Social Actors at RO-MAN (2008)
76. Lohse, M., Rohlfing, K.J., Wrede, B., Sagerer, G.: Try something else! when users change their discursive behavior in human-robot interaction. In: International Conference on Robotics and Automation (2008)
77. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int J Computer Vis **60**(2), 91–110 (2004)
78. Lucey, S., Ashraf, A.B., Cohn, J.F.: Face Recognition, chap. Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, pp. 275–286. I-TECH Education and Publishing (2007)
79. Ma, Y., Konishi, Y., Kinoshita, K., Lao, S., Kawade, M.: Sparse bayesian regression for head pose estimation. In: International Conference on Pattern Recognition, pp. 507–510 (2006)
80. Matthews, I., Baker, S.: Active appearance models revisited. Int J Computer Vis **60**, 135–164 (2004)
81. Maynard-Smith, J., Harper, D.: Animal Signals. Oxford University Press (2004)
82. McAndrew, F.T.: A cross-cultural study of recognition thresholds for facial expressions of emotion. J Cross-Cultural Psychol **17**, 211–224 (1986)
83. McLachlan, J.F.C.: A short adjective check list for the evaluation of anxiety and depression. J Clin Psychol **32**(1), 195–197 (1976)
84. McNaira, D.M., Lorr, M.: An analysis of mood in neurotics. J Abnorm Soc Psychol **69**(6), 620–627 (1964)
85. Mehrabian, A., Friar, J.T.: Encoding of attitude by a seated communicator via posture and position cues. J Consult Clin Psychol **33**(3), 330–336 (1969)
86. Mehrabian, A., Russell, J.A.: An Approach to Environmental Psychology. The MIT Press, Cambridge, MA, US (1974)
87. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Computer Vis Image Underst **98**(1), 4–24 (2005)
88. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. Pattern Analysis Machine Intelligence **31**(4), 607–626 (2009)
89. Newman, R., Matsumoto, Y., Rougeaux, S., Zelinsky, A.: Real-time stereo tracking for head pose and gaze estimation. In: International Conference on Automatic Face and Gesture Recognition, pp. 122–128 (2000)
90. Niit, T., Valsiner, J.: Recognition of facial expressions: An experimental investigation of ekman's model. Ada et Commentationes Universitatis Tarvensis **429**, 85–107 (1977)
91. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: International Conference on Computer Vision, pp. 1–8 (2007)
92. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois Press, Urbana, USA (1957)
93. Osgood, C.E.: Dimensionality of the semantic space for communication via facial expressions. Scand J Psychol **7**(1), 1–30 (1966)
94. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. Pattern Analysis Machine Intelligence **22**(12), 1424–1445 (2000)
95. Parkinson, B.: Do facial movements express emotions or communicate motives? Pers Soc Psychol Rev **9**(4), 278–311 (2005)
96. Poggi, I., D'Errico, F., Vincze, L.: Types of nods. the polysemy of a social signal. In: International Conference on Language Resources and Evaluation, pp. 17–23 (2010)
97. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol **17**, 715–734 (2005)
98. Provine, R.R., Fischer, K.R.: Laughing, smiling, and talking: Relation to sleeping and social context in humans. Ethol **83**(4), 295–305 (1989)

99. Rabie, A., Lang, C., Hanheide, M., Castrillón-Santana, M., Sagerer, G.: Automatic initialization for facial analysis in interactive robotics. In: International Conference on Computer Vision Systems, pp. 517–526 (2008)
100. Rabie, A., Wrede, B., Vogt, T., Hanheide, M.: Evaluation and discussion of multi-modal emotion recognition. In: International Conference on Computer and Electrical Engineering, vol. 1, pp. 598–602 (2009)
101. Ricciardelli, P., Baylis, G., Driver, J.: The positive and negative of human expertise in gaze perception. Cogn **77**(1), B1–B14 (2000)
102. Russell, J.A.: Evidence of convergent validity on the dimensions of affect. J Cross-Cultural Psychol **36**(10), 1152–1168 (1978)
103. Russell, J.A.: Affective space is bipolar. J Pers Soc Psychol **37**(3), 345–356 (1979)
104. Russell, J.A.: A circumplex model of affect. J Pers Soc Psychol **39**(6), 1161–1178 (1980)
105. Russell, J.A.: Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. Psychol Bull **115**(1), 102–141 (1994)
106. Russell, J.A.: Facial expressions of emotion: what lies beyond minimal universality? Psychol Bull **118**(3), 379–91 (1995)
107. Sakoe H.; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. Acoust Speech Signal Process **26**(1), 43–49 (1978)
108. Sander, D., Grandjean, D., Kaiser, S., Wehrle, T., Scherer, K.R.: Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. Eur J Cogn Psychol **19**(3), 470–480 (2007)
109. Sebe, N., Cohen, I., Gevers, T., Huang, T.S.: Emotion recognition based on joint visual and audio cues. In: International Conference on Pattern Recognition, vol. 1, pp. 1136–1139 (2006)
110. Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S.: Authentic facial expression analysis. Image Vis Comput **25**(12), 1856–1863 (2007)
111. Shacham, S., Dar, R., Cleeland, C.S.: The relationship of mood state to the severity of clinical pain. Pain **18**(2), 187–197 (1984)
112. Snider, J.G., Osgood, C.E. (eds.): Semantic differential technique. Aldine, Chicago (1969)
113. Storti, C.: Speaking of India: Bridging the Communication Gap When Working With Indians, chap. Yes, No, and other Problems, pp. 35–76. Nicholas Brealey Publishing (2007)
114. Thayer, R.E.: Measurement of activation through self-report. Psychol Rep **20**(2), 663–678 (1967)
115. Tian, Y.l., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. Pattern Analysis Machine Intelligence **23**(2), 97–115 (2001)
116. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. J Machine Learn Res **1**, 211–244 (2001)
117. Valstar, M.F., Gunes, H., Pantic, M.: How to distinguish posed from spontaneous smiles using geometric features. In: International Conference on Multimodal Interfaces, pp. 38–45 (2007)
118. Varchmin, A.C., Rae, R., Ritter, H.: Image based recognition of gaze direction using adaptive methods. In: Gesture and sign language in human-computer interaction. International Gesture Workshop. Bielefeld, Germany (1997)
119. Wang, J.G., Sung, E.: Study on eye gaze estimation. Syst Man Cybern **32**(3), 332–350 (2002)
120. Wang, J.G., Sung, E.: Em enhancement of 3d head pose estimated by point at infinity. Image Vis Comput **25**, 1864–1874 (2007)
121. Watson, O.M.: Proxemic Behavior: a Cross-Cultural Study. Mouton De Gruyter (1970)
122. Westbrook, M.T.: Positive affect: A method of content analysis for verbal samples. J Consult Clin Psychol **44**(5), 715–719 (1976)
123. Widen, S.C., Russell, J.A.: Descriptive and prescriptive definitions of emotion. Emot Rev **2**(4), 377–378 (2010)
124. Yang, P., Liu, Q., Cui, X., Metaxas, D.N.: Facial expression recognition based on dynamic binary patterns. In: Conference on Computer Vision and Pattern Recognition (2008)
125. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. Multimed **8**(3), 500–508 (2006)
126. Yoo, D.H., Kim, J.H., Lee, B.R., Chung, M.J.: Non-contact eye gaze tracking system by mapping of corneal reflections. In: International Conference on Automatic Face and Gesture Recognition, pp. 94–99 (2002)
127. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Pattern Analysis Machine Intelligence **31**(1), 39–58 (2009)
128. Zhao, G., Chen, L., Song, J., Chen, G.: Large head movement tracking using sift-based registration. In: International Conference on Multimedia, pp. 807–810 (2007)

Christian Lang is a Ph.D. student at Bielefeld University, Faculty of Technology. In 2008, he joined the CoR-Lab Research Institute for Cognition and Robotics and the Applied Informatics Group at Bielefeld University. He studied informatics at Chemnitz University of Technology and Bielefeld University, where he received the diploma with honors in 2007. His current research interest is focused on facial communicative signals, human-robot interaction, computer vision, and pattern recognition.

Sven Wachsmuth is holding a lecturer position at Bielefeld University and is currently heading the Central Lab Facilities of the Center of Excellence Cognitive Interaction Technology (CITEC). He received the Ph.D. degree in computer science from Bielefeld University in 2001. In 2003, he spent a sabbatical year at the Computer Science Department of the University of Toronto. His research interests are in human-robot interaction, especially looking at high-level computer vision problems, and system integration and evaluation aspects. He has been a member of the RoboCup@Home organization committee 2010-2011 and is member of the technical committee in 2012.

Marc Hanheide is a lecturer in the School of Computer Science at the University of Lincoln, UK. He received the diploma in computer science from Bielefeld University, Germany, in 2001 and the Ph.D. degree (Dr.-Ing.) also in computer science also from Bielefeld University in 2006. In 2001, he joined the Applied Informatics Group at the Technical Faculty of Bielefeld University where he worked in the European Union IST project VAMPIRE. From 2006 to 2009 he held a position as a senior researcher in the Applied Computer Science Group as a PI in the EU cognitive robotics project COGNIRON. From 2009 until 2011 he was a research fellow at the School of Computer Science at the University of Birmingham, UK, working in the EU cognitive robotics project CogX. Marc Hanheide also is a PI in a number of projects funded by the CoR-Lab and the Cluster of Excellence CITEC, Bielefeld. In all his work he researches on autonomous robots, human-robot interaction, interaction-enabling technologies, and architectures for cognitive systems.

Heiko Wersing received the diploma in physics in 1996 from Bielefeld University, Germany. In 2000, he received his Ph.D. in science from the Faculty of Technology, Bielefeld University. In 2000, he became a member of the Future Technology Research Group of Honda R&D Europe, GmbH, Offenbach, Germany. Currently he holds a position as a chief scientist in the Honda Research Institute Europe, at Offenbach. Since 2007, he is also co-speaker of the graduate school of the CoR-Lab Research Institute for Cognition and Robotics, Bielefeld University. His current research interests include recurrent neural networks, models of perceptual grouping and feature binding, principles of sparse coding, biologically motivated object recognition and online learning.