

Facial Expression Reactions to Feedback in a Human-Computer Interaction—Does Gender Matter?

Stefanie Rukavina*, Sascha Gruss, Holger Hoffmann, Harald C. Traue

Section Medical Psychology, Ulm University, Ulm, Germany

Email: *stefanie.rukavina@uni-ulm.de

Received 13 February 2016; accepted 21 March 2016; published 24 March 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The present study investigated facial expressions as reactions to positive and negative feedback during a human-computer interaction. Thirty subjects interacted in natural language with a computer system while conducting a mental task, which was designed in the style of a Wizard of Oz scenario. During six different experimental sequences from a well-known memory paradigm (“Concentration”), the wizard produced different feedback as a manipulative tactic to alter the users’ reactions, e.g., praising the subject was assumed to exert a positive influence on the subjects’ moods. To analyze the impact of various feedback on the emotional reaction, facial expressions from all subjects were coded with the Facial Expression Coding System (FACES). In general, the outcomes yielded more negative facial expressions in response to positive feedback as compared with negative feedback. Furthermore, a gender difference was evident: women showed significantly more negative facial expressions in response to praise (“Your performance is improving”) than men. The study and results show that when considering future companion technologies, it is important to carefully investigate users’ reactions to feedback, since paradoxical reactions may ensue.

Keywords

Affective Computing, Facial Expression, FACES, Positive and Negative Feedback, Gender Difference

1. Introduction

Affective computing is described by Rosalind Picard as “computing that relates to, arises from or deliberately

*Corresponding author.

influences emotions” (Picard, 1995). The main focus in affective computing is to build computers or other digital devices that are able to improve human-computer interactions (HCI). This endeavor not only involves an improved usability of the technology itself but is also more supportive and sensitive to the users’ needs. However, one important condition of improved usability is for the digital system to be able to recognize and respond respectively to the emotional and motivational state of its user (Calvo & D’Mello, 2010). If a cognitive technical system is empathetic and adaptive to the needs and abilities of a human user, it could be defined as a companion technology (Traue et al., 2013; Wendemuth & Biundo, 2012).

Overcoming this challenge of recognizing the emotional and dispositional states of a user in a robust manner and with high recognition accuracy, human-computer interactions would thus achieve a higher degree of quality. It would be possible to use such companion technologies as supportive digital companions, e.g., for people with special demands such as elderly individuals or as cited in Walter et al. (2013): “its application potential ranges from novel individual operation assistants for the technical equipment to a new generation of versatile organization assistants and digital services and, finally, to innovative support systems, e.g., for patients in rehabilitation or people with limited cognitive abilities” (Walter et al., 2013). Companion technology goes beyond assistive technology if the recognition of users’ mental states is used to adapt to and support the users’ goals through meaningful feedback.

Independent of the research for robust emotion and affect recognition in affective computing, we believe that our knowledge with respect to providing effective feedback to a user in an HCI context is rather limited (see 4 Discussion) and should become another important issue in affective computing: What kind of strategies can the technological system use to alter the users’ emotions if something goes awry during the interaction? Is it accurate to assume that positive feedback will automatically elicit a positive emotion? To answer these questions we analyzed facial expressions of subjects in an HCI scenario as reactions to feedback given by a cognitive technical system – that is, a simulated companion system, designed in the form of a Wizard of Oz experiment.

For human-human interaction facial expressions are of particular importance since they convey information about the intention, motivation and emotion of the conversational partner, who is either speaking or listening. With reference to the concept of embodied communication, every listener is also a sender and vice versa (Storch & Tschacher, 2014), which corresponds to the closed loop concept. Alex Pentland describes facial expressions and body movements as so-called “honest signals”. These signals can be taken as an indicator to measure the quality of an interaction between humans (Pentland, 2010). To underscore the role of these signals, facial expressions are also among behavioral changes throughout an emotional reaction. To be sure, there are also palpable and measurable changes in physiology, e.g., heart rate, subjective experience and cognitive ratings as emotions are complex phenomena (Traue & Kessler, 2003), but by measuring facial expressions in the HCI the inference is to use them as honest signals and therefore as an indicator of the underlying emotion, the intention, as well as motivation of the users.

In order to measure emotional responses, the majority of researchers focused on the recognition of discrete emotions, placing emphasis on basic emotions (Ekman, 1992). According to Ekman’s theory, there are six different and distinguishable facial emotions that can be recognized in every culture since they “evolved through adaption to our surroundings” (Ekman & Cordaro, 2011). The automatic identification of these six emotions however can be very challenging: facial expressions can be presented in different types as macro, subtle and micro expressions. Full-blown macro expressions are observable in an unregulated manner, that is, when it is permissible to convey these emotions according to display rules (Ekman et al., 2013) since they have the longest presentation time of several seconds. These expressions should be easily detected, for instance, in the form of posed emotions (Dhall et al., 2012). Yet naturally occurring facial expressions are far more difficult to recognize because they are rarely presented as macro expressions. It is more often the case that one finds subtle expressions. These expressions are diminished in their intensity due to regulation processes and are primarily observed in the upper portion of the face (due to lower somatic sensory representation). Lastly, micro expressions are intensive expressions, albeit very brief (only several ms), before a regulation can take place and therefore often referred to as “bodily leakage” (Ekman, 2003). Both mentioned types of expressions are difficult to identify automatically, on account of their low intensity, ambiguous location on the facial plane and short duration.

These aforementioned reasons have primarily led to the development of FACS (facial action coding system) (Ekman & Friesen, 1978; Ekman et al., 2002), a method enabling trained and certified persons to objectively describe facial movements according to specific action units as presented in various intensities. While this method is often employed and yields both high validity as well as high interrater agreement, it is very costly in

view of the time required and the need for certified FACS coders. In compliance with FACS, the method FACES (Facial Expression Coding System) was developed to circumvent these costs and was therefore applied in this study, as per 2.1 (Kring & Sloan, 2007).

There is some consensus among emotion experts who maintain that an expressed emotion naturally leads to a facial expression, along with individual differences, e.g., inhibition of facial expressiveness (Traue & Deighton, 2000) or the aforementioned display rules (Ekman & Friesen, 2003). However, not every facial expression must be based on a context-specific emotional experience and thus the source can remain unknown. This is referred to as “Othello’s error” (Kreidler, 2004). To elaborate, “Othello killed Desdemona because he thought that her signs of fear were of a woman caught in a betrayal. She was afraid of being disbelieved. The fear of being disbelieved looked just like the fear of being caught. Fear is fear. You have to find out which it is” (Ekman, 2004). This indicates that a facial expression shows a high degree of correlation with the emotional experience under certain circumstances (Keltner & Ekman, 2000), yet the exact source of the emotion can remain uncertain or unknown, e.g., unconscious emotions (Berridge, 2004). This aspect will not be further discussed here because it is also important to know how to positively influence the interaction as opposed to pinpointing the exact source of a specific emotion during an HCI. One opportunity to influence an ongoing interaction could involve providing specific positive feedback concerning the performance or motivation. This method was used in the following study in which different emotional states according to the VAD (valence, arousal and dominance) space were induced.

2. Methods

For this study “EmoRec II” a naturalistic experiment was designed using a Wizard of Oz design (Bernsen et al., 1994; Kelley, 1984), based on a previous experiment known as “EmoRec I” (Walter et al., 2013). In the experiment the users believed they were communicating in natural language and interacting with a technical system without any experimental manipulation, although they were indeed interacting with the experimenter “wizard.” This method allows the investigation of naturally elicited behavior which is flexible enough to make the user believe they are engaging in natural communication, yet standardized enough by script and predefined responses to obtain multimodal data for research. The experimental task followed the rules and precept of the popular game, “Concentration” (Walter et al., 2013). It consisted of six experimental sequences (ES) (see [Figure 1](#)), which were on the verge of inducing different emotional states by eliciting similar emotions in users according to the dimensional concept of emotions, valence, arousal and dominance, originally based on Wundt’s dimensional description (Wundt, 1896) and further developed by Russel and Mehrabian (Mehrabian & Russell, 1974; Russel & Mehrabian, 1977).

One strategy in this study for inducing different valenced emotional states was to manipulate the task difficulty itself, such as using more cards or very similar card motifs to increase the difficulty, thereby increasing the negativity of valence. The second strategy for inducing negative emotional states was to change the interaction itself. For example, the wizard could use a “delay”, which caused a system delay in terms of displaying the card (the user has verbally ordered) after six seconds. Another interactional manipulation was the option “wrong card”. Here, the wizard opened a false card, e.g., instead of a card at the location E1 (coordination system), the wizard opened D1. The third method of inducing emotions was by means of feedback from the technical system. The variation of the system’s behavior in response to the subjects was implemented via natural spoken language. In this case, the user was informed about their performance, e.g., if it was declining or improving. The user was also asked to articulate more clearly. The different manipulations and feedback in the experimental sessions were scripted and, in addition, the sequence of the presented positive and negative feedback was randomized. At the end of the experiment every subject rated each ES according to Valence, Arousal and Dominance with the SAM-Rating (Self Assessment-Rating Scale) according to (Bradley & Lang, 1994) as a manipulation check (see Results). It measures directly the dimensions valence, arousal and dominance non-verbally with graphic depictions for each dimension (1 - 9 Likert Scale). As Bradley and colleagues report “SAM has been used effectively to measure emotional responses in a variety of situations, including reactions to pictures [...], sounds [...] and more” (Bradley & Lang, 1994). For an example the reader may refer to the (Bradley & Lang, 1994).

The subjects’ overt behaviors were recorded by video and audio throughout the whole experiment (see [Figure 2](#)). This allowed the subsequent coding and analyzing of facial responses following the systems’ feedback. All final facial expression annotations including other recorded data (e.g., physiology) resulted in a data corpus

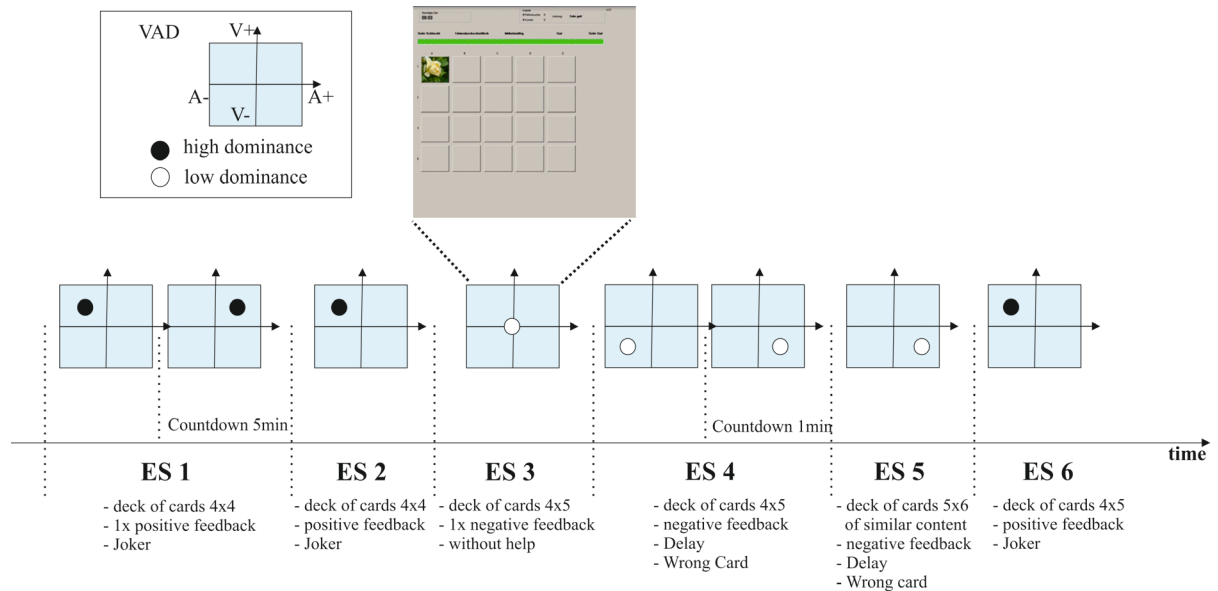


Figure 1. Experimental Design with six different experimental sequences (ES): Each of them was used to induce different emotional states according to the dimensional concept (valence, arousal, dominance).

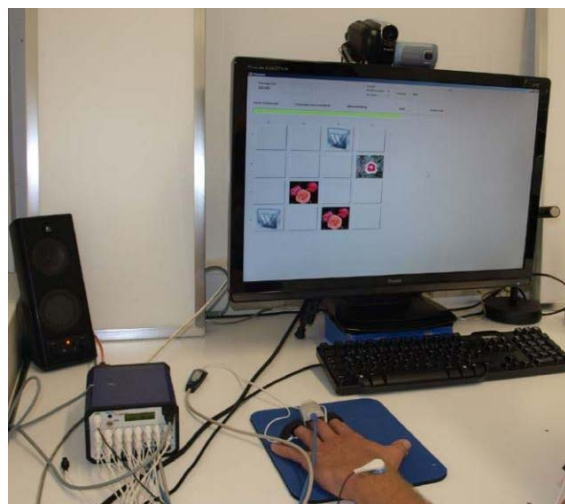


Figure 2. Experimental Setup: The subject interacted verbally with the technical system. In addition to the recording of physiological signals, the video and audio signal was also saved.

known as OPEN_EmoRec_II¹ (Rukavina, et al., 2015).

2.1. Facial Expression Coding System

For the analysis of facial expressions the method of Facial Expression Coding system (FACES) was used (Kring & Sloan, 2007), based on the more popular method FACS (Ekman & Friesen, 1978; Ekman, et al., 2002). FACES is an alternative option for coding facial expressions in specific situations. The main difference between both coding systems is that in FACES the individual muscle movements (action units) are not actually coded, rather, only the changes in facial expressions in general. Therefore this method offers a highly valid measure-

¹OPEN EmoRec II is a subsample (N = 30) of the total sample of this study. The 30 subjects gave permission to use their raw facial data from video and audio registration and to share this raw data set with other researchers for scientific purposes.

ment of observable facial reactions without the need for certified FACS coders and it is less time-consuming.

Every coder rates valence and intensity (Likert Scale 1 - 4) of each facial expression, thus the dimensional aspect of these emotional expressions can be advantageous if you induce emotions in a dimensional matter. In addition, every rater included the information of the subject's overall expressiveness (1 - 5 Likert scale) and noted the start time, the apex and end time of the emotional expression from which the duration could be calculated. To check for validity and agreement between different raters the intraclass correlation coefficient according to (Hallgren, 2012) was calculated, (see 3.2).

All verbal technical feedback was brief with a mean duration of 2.1 seconds. The delay system feedback took 6 seconds and to prevent overlaps between different feedback, the videos were cut after 6 seconds. The extracted videos (see **Table 1**) were numbered randomly and presented to the coders (without audio), thus they were taken out of context so that no numerical order correlations between different subjects could be established. All videos were presented without audio, and only visible facial expressions were used for coding.

2.2. Subjects

The sample consisted of 30 right handed subjects ($n = 23$ women; mean age: 37.5 years (± 19.5); $n = 7$ men; mean age: 51.1 years (± 26.5)). Every subject signed the informed consent form and thereby allowed the publication of their video, audio, and physiological signals. Every subject earned €35 for their participation. The study was conducted according to the ethical guidelines of Helsinki (certified by the Ulm University ethics committee: C4 245/08-UBB/se). In addition, the subjects agreed to the sharing of raw video data).

3. Results

3.1. Manipulation Check

Self Assessment Manikins largely reflected the hypothesized and expected ratings for every experimental sequence (see **Table 2**). We garnered high valence ratings for the sequences ES 1, ES 2 and ES 6. ES 4 and ES 5 induced negative emotions indicated by low valence ratings. ES 3 was designed to induce a neutral state. A Kruskal-Wallis and post hoc Bonferroni corrected tests showed significant differences in valence, arousal and dominance between ES 4 and ES 5 for all other sequences ($p < 0.00$).

Table 1. Quantity of extracted videos according to their feedback specificity: Performance-related (audio feedback) or system-related.

Feedback	n Videos
Negative feedback	208
Positive feedback	272
Delay	119
Wrong Card	67

Table 2. SAM Rating (mean and standard deviation).

Experimental sequence and aim of emotional induction (VAD-space)	Mean valence (SD)	Mean arousal (SD)	Mean dominance (SD)
ES 1 (+++)	7.70 (1.75)	3.88 (2.31)	7.59 (1.74)
ES 2 (+-+)	7.36 (1.84)	4.16 (2.16)	7.23 (1.78)
ES 3 (0-+)	6.96 (1.84)	4.71 (2.12)	6.67 (1.77)
ES 4 (-+-)	5.46 (2.38)	5.54 (2.07)	5.74 (2.16)
ES 5 (-+-)	3.91 (2.71)	6.42 (2.31)	4.19 (2.56)
ES 6 (+-+)	7.99 (1.51)	3.87 (2.41)	7.69 (1.64)

3.2. FACES

Although facial expressions occurred at various intervals in the course of the experiment, only facial reactions as direct responses to positive and negative feedback or system-related reactions such as delay and wrong card were selected for detailed analysis of the facial movement.

Criteria for the selection decision demanded that at least two raters must have detected and rated the facial expression, otherwise the expression was not taken into account. If there was an agreement, the event was labeled according to its valence and to the number of raters, which can be included in terms of agreement and therefore as a criterion of validity (see **Table 3**).

The ICC (3, 4) model (intraclass correlation coefficient, two-way randomized) was used to calculate the agreement between the four raters according to Hallgren resulting in an ICC = 0.74 (Hallgren, 2012). This value is the mean of the ICC for valence = 0.75, ICC for intensity = 0.65, ICC for duration = 0.70 and ICC for expressiveness = 0.84. It is evident that the four raters had a relatively high agreement on their ratings.

To calculate the type of valenced facial expression detected following the different feedback, all feedback that was presented less than 30 times in total for all subjects in the experiment were excluded from further analyses (see **Table 4**). Afterwards, we standardized the different facial expressions (VN, VP, VA, and no expression) on an individual basis. For each feedback we calculated the standardized expression as follows:

$$\text{standardized expression} = \frac{\text{n emotional facial expression/n expressions}}{\text{n feedback}}$$

with “emotional facial expression” = VN/VP/VA/no expression and “n expression” = sum of (VN + VP + VA) for specific feedback

Although we constructed the ambivalent label class to include facial expression detected by the coders with opposing valence, we did not further analyze these expressions because we believe it is a priority at this time to initially focus on definite valenced expressions in an ongoing human-computer interaction.

Figure 3 and **Figure 4** illustrate the standardized facial expression reactions to negative and positive feedback. The standard deviation is high due to high individual differences.

In general, the Wilcoxon test reveals that significant more facial expressions (independent of its valence) are shown in reaction to negative feedback ($U = -2.24$; $p = 0.025$). Furthermore, there can be significantly more negative expressions (mean = 0.0512) detected in response to positive feedback ($U = -2.849$; $p = 0.004$) compared to negative expressions in response to negative feedback (mean = 0.0328). No significance was found for the comparison of positive expressions after positive or negative feedback ($U = -0.487$; $p = 0.626$).

Table 3. FACES Label Class (VN = valence negative; VP = valence positive; VA= valence ambivalent; Due to the fact that the raters were not sure about the valence in some expressions the label “ambivalent” was added.

Label class	Description	n
VN_1	4 raters detected expression and agreed in valence	78
VN_0.75	3 raters detected expression and agreed in valence	97
VN_0.5	2 raters detected expression and agreed in valence	167
VP_1	4 raters detected expression and agreed in valence	41
VP_0.75	3 raters detected expression and agreed in valence	26
VP_0.5	2 raters detected expression and agreed in valence	44
VA_04	4 raters detected expression; 2 agreed in positive and 2 in negative	102
VA_03	3 raters detected expression; 2 agreed in valence	116
VA_02	2 raters detected expression; 1 rated it as positive and 1 as negative	86
Total		757

Next, we conducted several tests to detect gender differences. At first, we tested the rated expressivity for all subjects. We found no significant difference between male and female subjects ($U = -0.07$; $p = 0.96$; mean expressivity females = 2.26; males = 2.32).

Subsequently, a U-Test was conducted to look for gender effects in relation to the different types of feedback (see **Figure 5** and **Table 5**). Females tend to show significantly more negative expressions after the positive feedback, “Your performance is improving”. The effect size for this difference, according to Hedges (on account of different sample sizes) is $g = 0.91$ or according to Cohen’s $d = 0.98$, which can be considered a strong effect (Cohen, 2013).

For all other positive feedback and negative feedback, no gender difference could be found and further reports are therefore not included.

Table 4. Overview of feedback (quantity) and standardized facial expression.

Feedback (quantity)	Standardized negative expression*	Standardized positive expression*	Standardized ambivalent expression	Standardized no expression
neg1 (3)	0.00	0.50	0.50	0.00
neg2 (6)	0.12	0.12	0.26	0.00
neg3 (6)	0.73	0.00	0.10	0.00
neg4 (16)	0.42	0.06	0.44	0.00
neg5 (18)	0.46	0.25	0.29	0.00
neg6 (88)*	0.17	0.06	0.14	0.10
neg7 (71)*	0.16	0.14	0.22	0.11
neg8 (67)*	0.14	0.01	0.12	0.15
neg9 (119)*	0.16	0.06	0.26	0.11
pos1 (139)*	0.09	0.03	0.10	0.19
pos2 (46)*	0.38	0.10	0.26	0.13
pos3 (33)*	0.49	0.02	0.34	0.07
pos4 (28)	0.48	0.07	0.46	0.00
pos5 (26)	0.53	0.06	0.20	0.08

Notification: *only these are included for analysis. neg1 “I did not understand you, I am sorry”, neg2 “I did not understand you”, neg3 “I did not understand you, could you please repeat it?”, neg4 “Your articulation is unclear”, neg5 “Please articulate more clearly,” neg6 “Your performance is declining,” neg7 “Would you like to terminate the task?” neg8 “Delay”, neg9 “Wrong Card”, pos1 “Your performance is improving”, pos2 “Keep it up”, pos3 “You are doing great”, pos4 “Great”, pos5 “Your memory works perfectly.”) [All feedback was originally presented in German].

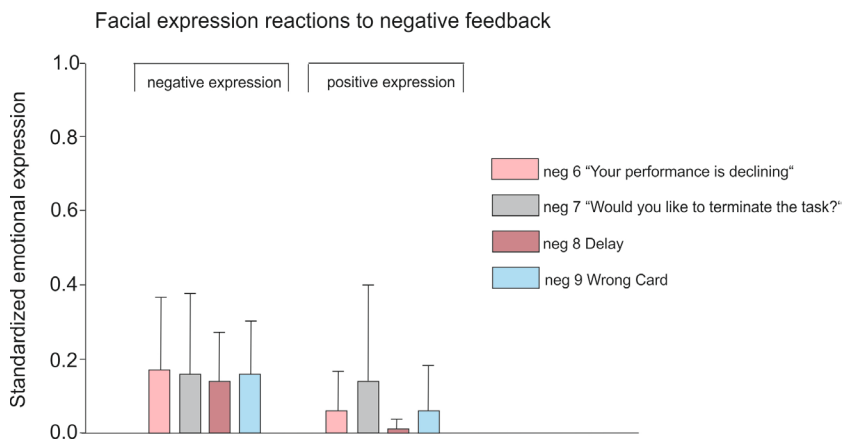


Figure 3. Standardized scores of facial expression reactions to negative feedback.

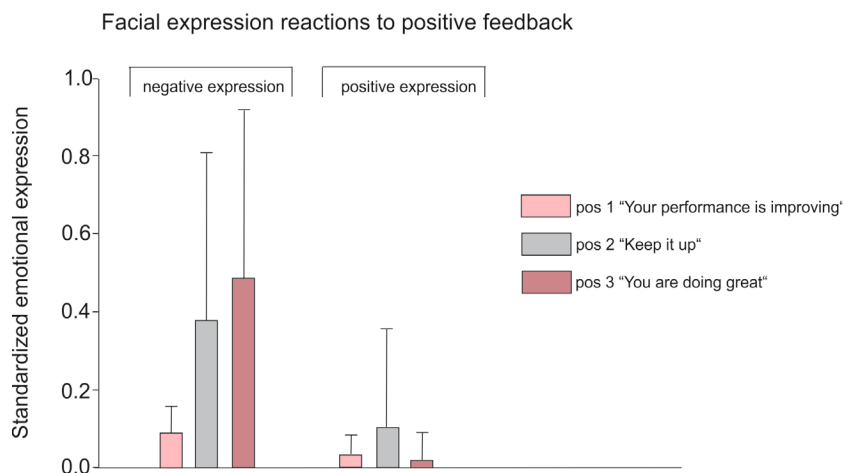


Figure 4. Standardized scores of facial expression reactions to positive feedback.

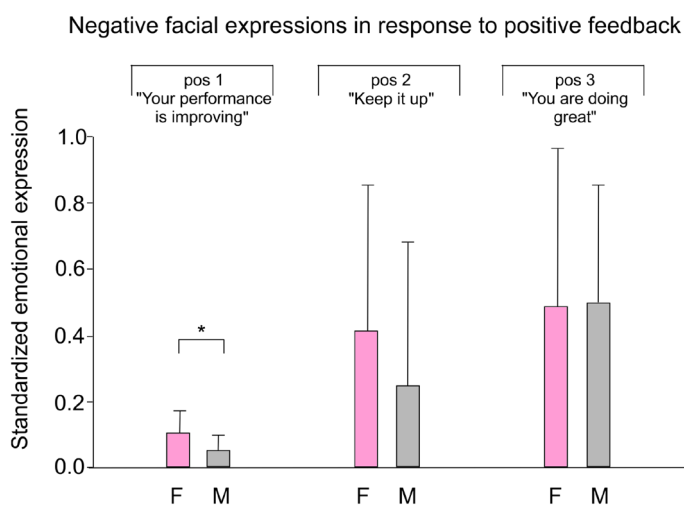


Figure 5. Gender differences within negative facial expressions in response to positive feedback (F = Females; M = Males).

Table 5. Impact of gender on facial expression reactions (positive or negative): standardized scores of facial expressions after specific positive feedback (pos1 "Your performance is improving", pos2 "Keep it up", pos3 "You are doing great").

Gender	Positive facial expression			Negative facial expression		
	pos1	pos2	pos3	pos1	pos2	pos3
Female:mean (sd)	0.03 (0.06)	0.05 (0.16)	0.01 (0.05)	0.11 (0.07)	0.42 (0.44)	0.49 (0.48)
Male:mean (sd)	0.02 (0.02)	0.30 (0.45)	0.05 (0.11)	0.05 (0.05)	0.25 (0.43)	0.50 (0.35)
U	-0.13	-1.35	-0.75	-2.05	-0.90	-0.16
p	0.92	0.37	0.72	0.04	0.41	0.87

4. Discussion

Affective computing aims at making technologies more supportive and empathic. Being able to respond to users' mental states offers many opportunities for digital devices and applications as Janssen asserts: "Although there are many challenges ahead, the opportunities for and promises of incorporating empathy into affective computing and social signal processing are manifold. When such research comes to fruition, it can enhance empathy,

thereby boosting altruism, trust, and cooperation. Ultimately, this could improve our health and well-being and greatly improve our future societies” (Janssen, 2012). Although there is extensive research concerning the best way to recognize and classify emotions robustly (Calvo & D’Mello, 2010), it still remains unclear what options are available for technology in order to intervene during an ongoing interaction. There is not much literature relating to affective interventions during a human-computer interaction. Klein et al. found that people who were frustrated by a computer would interact longer with the same computer after an intervention by an “affect-support agent” who showed empathy and sympathy (Klein et al., 2002). Hoques and colleagues found higher smile activity when using automatic video analysis (Hoque et al., 2012) while Partala and Surakka detected that people smiled more in stressful situations (Partala & Surakka, 2004).

There are several possibilities for digital technologies to recognize human emotional responses since emotions are multidimensional phenomena (Traue & Kessler, 2003). Facial expressions are counted among the so-called honest signals during human-human interactions (Pentland, 2008). In the past, research concerning affective computing focused especially on emotions according to Ekman’s suggested basic emotions, e.g., happiness and sadness (Ekman, 1992). Although several prototypical corpora for facial expressions (Kanade et al., 2000) have been developed, these databases predominantly consist of posed facial expressions. These lead to high intensity facial expressions or macro expressions, which are easy to detect even with algorithms on video data. To detect emotional expressions in a naturally occurring dialogue, whether between humans or between human and technical surfaces involving different dynamics and intensities, is much more challenging. In this study we decided to annotate the facial expressiveness utilizing Kring’s et al. FACES (Kring & Sloan, 2007). All 30 subjects, belonging to the OPEN_EmoRec_II corpus (Rukavina et al., 2015), were labeled by four coders who achieved an ICC of 0.74, thus guaranteeing a high agreement between the raters. Additionally, the final expressions were only labeled on the condition that at least two of the raters had seen this expression (see methods). After each experiment a SAM-Rating (Bradley & Lang, 1994) was collected for each experimental sequence as a manipulation check, thus we can conclude that the general induction of negative and positive emotional states with the help of different technical feedback was successful.

Facial expressions are considered to carry “aspects [...] of an emotional response and of social communication” (Adolphs, 2002), enabling us to draw conclusions about the human’s emotion or motivation. If a facial expression is presented as a communicational signal, which of course is assumed in an interaction and recognized by the user being videotaped, then it can be concluded that these visible facial expressions are known to be seen in the face and are intended to convey information about the internal emotional state of the user. This is one reason we found it noteworthy to analyze facial expressions with FACES according to (Kring & Sloan, 2007), which has two important advantages compared to the more popular method FACS according to Ekman: lower expenditure of time and no requirement for certified coders (Ekman & Friesen, 1978).

Comparing the facial expressions in this study to different technological feedback (positive and negative), revealed three interesting aspects. In general, a larger amount of negative expressions after positive feedback (see **Figure 3**) could be found. Although such a finding seems to be paradoxical, it corresponds to former research on feedback in a teacher and student situation and therefore is not reported for the first time. Meyer concludes that “praise can lead the recipient to infer that the other person evaluates his or her ability as low, while criticism can lead the recipient to conclude that his or her ability was estimated as high” (Meyer, 1992).

A second interesting fact is that there seems to be a noticeable gender difference. In our study women more often reveal significantly negative facial expressions in response to the positive feedback “Your performance is improving.” This could be explained by a combination of the paradoxical finding of praise in general and that in particular women tend to attribute success externally during a human-computer interaction (Dickhäuser, 2001). Or to be more precise, women attribute their own success as a coincidence which is the reason they judge the positive feedback to actually be more “negative”.

Another striking result is the high number of positive expressions upon receiving the critical, if not negative feedback, “Would you like to terminate the task?” This feedback was believed to induce negative emotions, which is why it was only used in ES 5. The explanation for this finding could be twofold: either the subjects tend to react in a positive manner because they feel liberated from the difficult situation, or the subjects tend to be embarrassed, which is often accompanied by a smile (Keltner & Anderson, 2000). In addition to these findings, Hoques and colleagues showed that users habitually smile more often in stressful situations (Hoque & Picard, 2011), in accordance with findings by Partala and Surakka who report a higher Zygomaticus major muscle activity measured by an electromyogram upon intervention in the face of negative feedback (Partala & Surakka,

2004).

To conclude the results of our study, we found that feedback of different valence during an ongoing human-computer interaction can have paradoxical consequences. It is thus erroneous to expect positive emotions (measured by facial expressions) after positive feedback. This context specificity seems to be further influenced by gender since women tend to show significantly more negative expressions.

5. Limitations and Outlook

The main limitation of the corpus OPEN_EmoRec_II is its sample size of 30 subjects (Rukavina, et al., 2015). The sample is low on account of the small number of subjects who consented to publishing their data for research. It would have been interesting to look for gender differences in a gender and age balanced population. It is also of significance that the subjects' personalities influenced emotional expressiveness as Limbrecht-Ecklundt and her peers discovered in her study of emotional expressiveness using the facial action coding system (Limbrecht-Ecklundt et al., 2013).

Concerns regarding the analysis of facial expressions itself relating to the subjective emotional experience is often debated, "as we do not know what the exact internal and external conditions are" (Kappas, 2003). This is a well-known problem and should be avoided or reduced in future devices measuring emotional states by using content-specific information and multimodal signals, e.g., physiology and audio combined with video signals (Caridakis et al., 2007; Walter et al., 2011). However, the authors would also like to point out that for this study we do not claim to know what the subjects were feeling in these situations. Rather, the SAM ratings for each experimental sequence show positive valence ratings at least for ES 1, ES 2 and ES 6, assuming the subjects liked the task independent of the higher rates of negative facial expressions. This automatically highlights the need for multimodal emotion classification since facial expressions as the sole basis are not always indicative of the truth that is important for the task at hand.

For subsequent research the authors suggest including physiology to look for further differences between positive and negative feedback or to explore feedback-specific physiological reactions. Pairing these results with the analysis of facial expressions could lead to varied or more specific results. Furthermore the labeling of facial expressions with the Facial Action Coding System should be considered in the future as this method is more likely to show differences on a finer scale and specific to discrete basic emotions (Ekman & Friesen, 1978, 1985). One possibility could be the usage of CERT (The Computer Expression Recognition Toolbox) software for detecting 19 AUs (Littlewort et al., 2011) which has already been proven successful in achieving favorable classification results (Bosch et al., 2014). This must be tested using the data available in this corpus since the subjects do have electrodes on two different sites. For more information see OPEN_EmoRec_II (Rukavina et al., 2015).

Acknowledgements

This research was supported in part by grants from the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). We also would like to thank Nike Schier and Gerrit Blum for analyzing the annotation results.

References

- Adolphs, R. (2002). Recognizing Emotion from Facial Expressions: Psychological and Neurological Mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1, 21-62. <http://dx.doi.org/10.1177/1534582302001001003>
- Bernsen, N. O., Dybkjær, H., & Dybkjær, L. (1994). Wizard of oz Prototyping: How and When. Paper Presented at the CCI Working Papers in Cognitive Science and HCI, Roskilde, Denmark.
- Berridge, K. C. (2004). Pleasure, Unfelt Affect, and Irrational Desire. In A. S. R. Manstead, N. H. Frijda, & A. E. Fischer (Eds.), *Feelings and Emotions: The Amsterdam Symposium Cambridge*. Cambridge, UK: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511806582>
- Bosch, N., Chen, Y., & D'Mello, S. (2014). It's Written on Your Face: Detecting Affective States from Facial Expressions While Learning Computer Programming. Paper Presented at the Intelligent Tutoring Systems. http://dx.doi.org/10.1007/978-3-319-07221-0_5
- Bradley, M. M., & Lang, P. J. (1994). Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, 49-59. [http://dx.doi.org/10.1016/0005-7916\(94\)90063-9](http://dx.doi.org/10.1016/0005-7916(94)90063-9)

- Calvo, R., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing, 1*, 18-37. <http://dx.doi.org/10.1109/T-AFFC.2010.1>
- Caridakis, G., Castellano, G., Kessous, L., Raouzaoui, A., Malatesta, L., Asteriadis, S. et al. (2007). Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech. In C. Boukis, A. Pnevmatikakis, & L. Polymenakos (Eds.), *Artificial Intelligence and Innovations 2007: From Theory to Applications*. Boston: Springer.
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *MultiMedia, IEEE, 19*, 34-41. <http://dx.doi.org/10.1109/MMUL.2012.26>
- Dickhäuser, O. (2001). *Computernutzung und Geschlecht*. Münster und New York: Waxmann Verlag.
- Ekman, P. (1992). Are There Basic Emotions? *Psychological Review, 99*, 550-553.
- Ekman, P. (2003). Darwin, Deception, and Facial Expression. *Annals of the New York Academy of Sciences, 1000*, 205-221. <http://dx.doi.org/10.1196/annals.1280.010>
- Ekman, P. (2004). Face to Face—The Science of Reading Faces. <http://globetrotter.berkeley.edu/people4/Ekman/ekman-con3.html>
- Ekman, P., & Cordaro, D. (2011). What Is Meant by Calling Emotions Basic. *Emotion Review, 3*, 2364-2370. <http://dx.doi.org/10.1177/1754073911410740>
- Ekman, P., & Friesen, W. V. (1978). *Manual for the Facial Action Code*. Palo Alto: Consulting Psychologist Press.
- Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues*. Los Altos, CA: Malor Books.
- Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). Can Facial Behavior Be Controlled Or Disguised? In P. Ekman, W. V. Friesen, & P. Ellsworth (Eds.), *Emotion in the Human Face—Guidelines for Research and an Integration of Findings*. Fairview Park: Pergamon Press Inc.
- Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System. Manual and Investigator's Guide*. Salt Lake City, UT: Research Nexus.
- Friesen, W. V., & Ekman, P. (1985) The Emotion Facial Action Coding System (EMFACS). Unpublished Document.
- Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23-34.
- Hoque, M. E., McDuff, D. J., & Picard, R. W. (2012). Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. *IEEE Transactions on Affective Computing, 3*, 323-334. <http://dx.doi.org/10.1109/T-AFFC.2012.11>
- Hoque, M., & Picard, R. W. (2011). Acted vs. Natural Frustration and Delight: Many People Smile in Natural Frustration. *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, Santa Barbara, 21-25 Mar 2011, 354-359.
- Janssen, J. H. (2012). A Three-Component Framework for Empathic Technologies to Augment Human Interaction. *Journal on Multimodal User Interfaces, 6*, 143-161. <http://dx.doi.org/10.1007/s12193-012-0097-5>
- Kanade, T., Cohn, J. F., & Tian, Y. L. (2000). Comprehensive Database for Facial Expression Analysis. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, 28-30 Mar 2000, 46-53. <http://dx.doi.org/10.1109/afgr.2000.840611>
- Kappas, A. (2003). *What Facial Activity Can and Cannot Tell Us about Emotions*. In M. Katsikitis (Ed.), *The Human Face* (pp. 215-234). New York: Springer. http://dx.doi.org/10.1007/978-1-4615-1063-5_11
- Kelley, J. F. (1984). An Iterative Design Methodology for User-Friendly Natural-Language Office Information Applications. *ACM Transactions on Office Information Systems, 2*, 26-41. <http://dx.doi.org/10.1145/357417.357420>
- Keltner, D., & Anderson, C. (2000). Saving Face for Darwin the Functions and Uses of Embarrassment. *Current Directions in Psychological Science, 9*, 187-192. <http://dx.doi.org/10.1111/1467-8721.00091>
- Keltner, D., & Ekman, P. (2000). Facial Expression of Emotion. In M. Lewis, & J. Haviland-Jones (Eds.), *Handbook of Emotions* (2nd ed., pp. 236-249). New York: Guilford Publications, Inc.
- Klein, J., Moon, Y., & Picard, R. W. (2002). This Computer Responds to User Frustration: Theory, Design, and Results. *Interacting with Computers, 14*, 119-140. <http://dx.doi.org/10.1111/1467-8721.00091>
- Kreisler, H. (2004). *Paul Ekman Interview: Conversations with History*. Berkeley, CA: Institute of International Studies, UC Berkeley. <http://globetrotter.berkeley.edu/people4/Ekman/ekman-con3.html>
- Kring, A., & Sloan, D. (2007). The Facial Expression Coding System (FACES): Development, Validation, and Utility. *Psychological Assessment, 19*, 210-224. <http://dx.doi.org/10.1037/1040-3590.19.2.210>
- Limbrecht-Ecklundt, K., Rukavina, S., Walter, S., Scheck, A., Hrabal, D., Tan, J.-W., et al. (2013). The Importance of Subtle Facial Expressions for Emotion Classification in Human-Computer Interaction. In A. Freitas-Magalhães (Ed.), *Emotional*

- Expression: The Brain and the Face* (Vol. 5). Porto: UFT Press.
- Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., et al. (2011). The Computer Expression Recognition Toolbox (CERT). *IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, 21-25 March 2011, 298-305. <http://dx.doi.org/10.1109/fg.2011.5771414>
- Mehrabian, A., & Russell, J. A. (1974). *An Approach to Environmental Psychology*. Cambridge, MA: The MIT Press.
- Meyer, W.-U. (1992). Paradoxical Effects of Praise and Criticism on Perceived Ability. *European Review of Social Psychology*, 3, 259-283. <http://dx.doi.org/10.1080/14792779243000087>
- Partala, T., & Surakka, V. (2004). The Effects of Affective Interventions in Human-Computer Interaction. *Interacting with Computers*, 16, 295-309. <http://dx.doi.org/10.1016/j.intcom.2003.12.001>
- Pentland, A. S. (2008). *Honest Signals: How They Shape Our World*. Cambridge, MA, London, England: MIT Press.
- Pentland, A. S. (2010). To Signal Is Human. *American Scientist*, 98, 204-219.
- Picard, R. W. (1995). Affective Computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321.
- Rukavina, S., Gruss, S., Walter, S., Hoffmann, H., & Traue, H. C. (2015). OPEN_EmoRec_II—A Multimodal Corpus of Human-Computer Interaction. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9, 977-983.
- Russel, J. A., & Mehrabian, A. (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11, 273-294. [http://dx.doi.org/10.1016/0092-6566\(77\)90037-X](http://dx.doi.org/10.1016/0092-6566(77)90037-X)
- Storch, M., & Tschacher, W. (2014). *Embodied Communication: Kommunikation beginnt im Körper, nicht im Kopf*. Bern: Verlag Hans Huber.
- Traue, H. C., & Deighton, R. W. (2000). Emotional Inhibition. In G. Fink (Ed.), *Encyclopedia of Stress*. San Diego, CA: Academic Press.
- Traue, H. C., & Kessler, H. (2003). Psychologische Emotionskonzepte. In A. Stephan, & H. Walter (Eds.), *Natur und Theorie der Emotion* (pp. 20-33). Paderborn: Mentis.
- Traue, H. C., Ohl, F., Brechman, A., Schwenker, F., Kessler, H., Limbrecht, K., et al. (2013). A Framework for Emotions and Dispositions in Man-Companion Interaction. In M. Rojc, & N. Campbell (Eds.), *Converbal Synchrony in Human-Machine Interaction* (pp. 99-140). Boca Raton, FL: CRC Press. <http://dx.doi.org/10.1201/b15477-6>
- Walter, S., Jonghwa, K., Hrabal, D., Crawcour, S. C., Kessler, H., & Traue, H. C. (2013). Transsituational Individual-Specific Biopsychological Classification of Emotions. *IEEE Transactions on Systems, Man, and Cybernetics*, 43, 988-995. <http://dx.doi.org/10.1109/TSMCA.2012.2216869>
- Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., et al. (2011). Multimodal Emotion Classification in Naturalistic User Behavior. *Proceedings of the 14th International Conference on Human-Computer Interaction: Towards Mobile and Intelligent Interaction Environments*, Part III, Orlando, 9-14 July 2011, 603-611. http://dx.doi.org/10.1007/978-3-642-21616-9_68
- Wendemuth, A., & Biundo, S. (2012). A Companion Technology for Cognitive Technical Systems. *COST 2102 International Training School*, Dresden, 21-26 February 2011, 89-103. http://dx.doi.org/10.1007/978-3-642-34584-5_7
- Wundt, W. M. (1896). *Lecture on Human and Animal Psychology*. New York: Macmillan.