Available online at www.sciencedirect.com
SCIENCE DIRECT®
Interacting with Computers xx (2005) 1–19
Interacting with Computers
www.elsevier.com/locate/intcom

Real-time estimation of emotional experiences from facial expressions

Timo Partala[a,*], Veikko Surakka[a,b], Toni Vanhala[a]

[a]*Tampere Unit for Computer–Human Interaction, Department of Computer Sciences, University of Tampere, FIN-33014 Tampere, Finland*
[b]*Department of Clinical Neurophysiology, Tampere University Hospital, P.O. Box 2000, FIN-33521 Tampere, Finland*

Received 18 October 2004; revised 6 May 2005; accepted 15 May 2005

Abstract

The present aim was to develop methods that estimate emotional experiences in real time from the electromyographic activity of two facial muscles: *zygomaticus major* (activated when smiling) and *corrugator supercilii* (activated when frowning). Ten subjects were stimulated with a series of emotionally arousing pictures and videos. After each stimulus the subjects rated the valence of their emotional experience on a nine-point bipolar dimensional scale. At the same time the computer estimated the subjects' ratings on the basis of their electrical facial activity during each stimulation with 70 computational models. The models estimated the subjects' ratings either categorically or dimensionally with regression models. The best categorical models were able to estimate negative and positive ratings with an average accuracy of over 70 and 80% for pictures and videos, respectively. The best correlations between the human ratings and machine estimations formed with the regression models were high ($r > 0.9$). These findings indicate that models estimating psycho-emotional experiences on the basis of facial activity can be created successfully in several ways.

Keywords: Emotions; Facial expression; Estimation; Human–computer interaction; Social agent; Psychophysiology

* Corresponding author. Tel.: +358 3 215 8555; fax: +358 3 215 6070.
E-mail address: tpa@cs.uta.fi (T. Partala).

0953-5438/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.intcom.2005.05.002

1. Introduction

Emotions can be defined as either discrete categories or continuous dimensions. A well-known discrete emotions framework is the argument for six basic emotions, that is anger, disgust, fear, joy, sadness, and surprise by Ekman (1993). On the other hand, according to the dimensional emotions framework, emotions are defined as a set of bipolar dimensions, which together define the emotion space (Bradley and Lang, 1994, 2000). Out of the original three dimensions (i.e. valence, arousal, and dominance), valence and arousal are the two most commonly used dimensions. The valence dimension varies from negative to positive emotional experience. The middle of the dimension represents a neutral emotional experience. The arousal dimension varies from calm to highly aroused. Again, the middle of the dimension represents a neutral experience.

Affective computing can be defined as computing that 'relates to, arises from or deliberately influences the user's emotions' (Picard, 1997). In affective computing, one way of estimating the user's affective state is by using physiological measurements. Examples of recent advances in affective computing include the emotion mouse, which measures the user's skin temperature, galvanic skin response (GSR), and heart rate (Ark et al., 1999). Another example is car drivers' stress recognition by measuring heart rate, activations of the *trapezius* muscle, respiration and galvanic skin response (Healey, 2000). It has also been suggested that pupil size and facial electromyography (EMG) could be potentially useful input signals in HCI (Partala and Surakka, 2003, 2004).

Facial expressions are an important source of emotion-related information. In human–human communication, facial emotional expressions can have a significant role (Surakka and Hietanen, 1998). In human–computer interaction, facial expression measurements could be used to provide information about the user's emotions without interrupting the user. By using that information, the computer could change its behavior according to the user's emotions and even show emotional intelligence skills similar to those of humans. Proposed areas that could utilize information provided by facial expressions include learning, entertainment and ubiquitous computing (Lisetti and Schiano, 2000). Even though facial expression measurement is regarded as a promising computer input method, existing systems utilizing information from the user's facial expressions are still rare. One example is expression glasses by Scheirer et al. (1999), which measured the activity of two facial muscles: *corrugator supercilii* and *frontalis*. Based on the information provided by the measurements, the system was able visualize the wearer's confusion and interest levels.

In the field of facial expression recognition, many efforts have been made in trying to recognize expressions of discrete emotions, especially the ones suggested by Ekman (1993). At present, different machine vision techniques using video cameras are the predominant methods in measuring facial expression (e.g. Cohen et al., 2003; Dailey et al., 2002; Oliver et al., 2000; Smith et al., 2001). In order to use facial expression measurements as an input signal in affective computing, it is important to know how facial expressions relate to the underlying emotional experiences. Consequently, the estimation of emotional experiences from objectively measured facial expressions becomes an important research topic.

Many current facial recognition systems rely on analyzing single facial images instead of tracking the changes in facial expressions continuously. However, Essa and Pentland

(1997) suggested that the lack of temporal information is a significant limitation in many facial expression recognition systems. In human–computer interaction, it would be important that the computer could analyze the user's facial expressions continuously to be able to react to changes in the user's emotional state at the right time. Thus, when developing methods for analyzing facial expressions in human–computer interaction, a real-time analysis is essential. This can be achieved either by using advanced video-based techniques (e.g. Essa and Pentland, 1997) or by measuring the electrical activity of muscles with EMG.

Research using a real-time analysis of facial expressions during human–computer interaction has not been very extensive. However, it is known that the users' facial expressions during computer usage are related to the different events in human–computer interaction. For example, Partala and Surakka (2004) found that computerized affective interventions affected the electrical activity of two facial muscles: *zygomaticus major* (the muscle that draws the lip corners up producing a smile) and *corrugator supercilii* (the muscle that knits and lowers the brows producing a frown), as well as experienced affective valences and cognitive performance.

Although real time analyses have been rare, the relationship of these two facial muscles and emotional experiences has been studied offline in many studies. There is evidence that the activations of *zygomaticus major* and *corrugator supercilii* are related to the experience of affective valence. Dimberg (1990) found increased *corrugator supercilii* EMG activity in response to angry facial stimuli, and increased *zygomaticus major* EMG activity in response to happy facial stimuli. Greenwald et al. (1989) found a significant negative linear covariation of *corrugator supercilii* EMG activity and ratings of experienced valence. Their data also showed the presence of a linear relation between *zygomaticus major* activity and ratings of valence, even though a quadratic relationship was stronger for this particular muscle. In addition, a score based on subtracting the *corrugator supercilii* EMG activity from *zygomaticus major* EMG activity showed a positive linear correlation and an even stronger quadratic trend.

Some of these results were replicated in Lang et al. (1993). Those results showed a negative linear trend between *corrugator supercilii* EMG activity and ratings of experienced valence, and a positive linear trend between *zygomaticus major* EMG activity and ratings of experienced valence. A negative linear relationship between *corrugator supercilii* EMG activity and ratings of experienced valence was also found using affective sounds as stimuli (Bradley and Lang, 2000). Recently, Larsen et al. (2003) found a negative linear correlation between *corrugator supecilii* EMG activity and valence ratings for affective pictures and sounds for female subjects. They also found a weaker linear correlation between *zygomaticus major* EMG activity and valence ratings.

Although there is evidence for universal facial expressions of certain emotions, it is important to realize that there are also differences in the facial behavior of different people. For example, Ekman (1985) stated that accurate interpretation of facial expression benefits from the knowledge of what is normative for the specific individual. In concordance with this, Cohn et al. (2002) found evidence for relatively stable individual differences in facial expressions. There was a correlation of 0.58 in *zygomaticus major* EMG responses to the same emotionally positive film clips between two sessions with a 12 months measurement interval. These findings have some implications for the design of emotion estimation from

facial expressions. First, the findings that there are significant differences in facial behavior between individuals suggest that the best results in emotion estimation could be obtained using a person-adaptive system, which forms an individual model of facial behavior for each individual user. This way the system would learn the facial expression style of each user. In addition, the previous findings suggest that once such a model has been formed, it can be used relatively successfully in subsequent computer use sessions, too.

The results from offline studies have been obtained by analyzing the results of large studies with lots of test subjects. In addition, the results have been obtained by analyzing averaged responses to many repetitions of affective stimuli. In human–computer interaction, an important question is how accurately emotions could be estimated from facial expressions, when the computer has to estimate the user's emotional experiences in real time from single responses instead of being able to analyze a large pool of data.

Based on the above findings we constructed a person-adaptive system, which estimated the subject's experienced emotions in real time from the activations of facial muscles by using predefined models. The estimation accuracies of 70 different computational models were tested in an experiment using this system. The purpose of the experiment was to study, how reliably experienced affective valence could be estimated in real time from the EMG measurements of the *zygomaticus major* and the *corrugator supercilii* facial muscles. The facial EMG responses were measured, while the test subjects were watching emotionally arousing pictures and videos.

2. Methods

2.1. Subjects

Ten technically successful recordings were made from five female and five male subjects (mean age 28.8 years, range 23–35 years). The subjects had normal or corrected-to-normal vision.

2.2. Overview

The structure of the experiment is shown in Fig. 1. First, there was a short practice phase, in which the subject just practiced giving the subjective ratings, and facial expression data was not measured. Second, there was a calibration phase, in which the subject's facial responses to 24 emotionally arousing pictures were first registered. The subjects also rated their experiences after each picture. After 24 pictures and ratings the adaptive estimation models were formed as the computer calculated 70 different models based on the subject's facial behavior and experience ratings. The model formation was invisible to the user. Third, the estimation accuracies of the models were tested as the subject saw 28 emotionally arousing pictures. After each picture, the system calculated 70 estimations (one for each model) of the subject's emotional experience based on the subject's facial behavior during the picture. Fourth, the estimation accuracies of the models were further tested as the subject saw six emotionally arousing video clips.

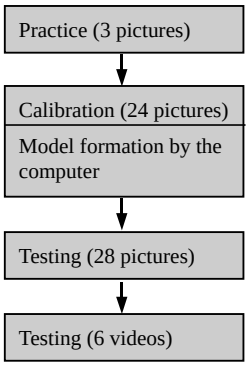

Fig. 1. Overview of the experimental setup.

Again, the system calculated 70 estimations of the subject's emotional experience of each video clip based on the subject's facial behavior during the clip.

2.3. Stimuli

2.3.1. Pictures

The picture stimuli were taken from the International Affective Picture System (IAPS) developed by Lang et al. (1995). Erotic stimuli were not used. The IAPS mean valence rating values reported in IADS ranging from 1 (negative) to 9 (positive) were used as criteria for stimulus selection.

For the calibration block, the stimuli were selected and divided into three categories, eight stimuli per category: negative stimuli (e.g. a burn victim and a corpse) with a mean valence of 2.0 and a mean arousal of 5.0, neutral stimuli (e.g. a towel and a mug) with a mean valence of 5.0 and a mean arousal of 2.0, and positive stimuli (e.g. a baby laughing and a beautiful scenery) with a mean valence of 8.0 and a mean arousal of 5.0. Strong negative and positive pictures were selected to ensure that the subjects' facial responses were large enough so that the estimation models could be successfully formed.

For the testing block, to facilitate regression analysis, the stimuli were selected so that they were evenly distributed on the valence dimension. Both high arousal and low arousal pictures were used so that the sample picture set used in the block imitated the distribution of the whole IAPS picture set on the arousal dimension (Lang et al., 1995).

2.3.2. Videos

Six video stimuli were used in the experiment. The video clips included facial expressions produced by a male and a female actor. Both actors acted a strong negative, neutral, and strong positive video clip (acting very sad, a neutral expression, and a happy laughter, respectively). All the video stimuli were digitized using the Adobe Premiere™ software, they were 30 s long, and they were soundless.

2.4. Models

2.4.1. Regression models

Twenty-five different regression-based models were used to estimate linearly the subjects' experienced affective valence based on facial EMG signals. For each model, a regression line, which minimizes the sum of squared errors, was calculated. The difference of the estimated valence rating and the actual valence rating was used as the error value in model formation. The outputs of these models were estimations of the subject's ratings of affective valence on a 1 (negative) –9 (positive) scale. The outputs were calculated using linear regression separately for each stimulus in the testing phases.

The 25 regression models were formed for the combinations of EMG signals and subjective valence ratings as follows:

1. The linear regression of *zygomaticus major* EMG and experienced valence ratings (regression line calculated with subjective ratings from 1 to 9).
2. The linear regression of *corrugator supercilii* EMG and experienced valence ratings (regression line calculated with subjective ratings from 1 to 9).
3. The linear regression of an EMG difference score (*zygomaticus major* EMG activity minus *corrugator supercilii* EMG activity) and experienced valence ratings (regression line calculated with subjective ratings from 1 to 9).
4. The linear regression of *zygomaticus major* EMG and experienced valence ratings (regression line calculated with subjective ratings from 5 to 9 only).
5. The linear regression of *corrugator supercilii* EMG and experienced valence ratings (regression line calculated with subjective ratings from 1 to 5 only).

The purpose of the last two models was to use only the part of the valence scale, which has the strongest correspondence for the activations for the particular muscle. For example, *zygomaticus major* is mostly activated in response to positive affect, and the positive end (5–9) was used in model formation.

In addition, all the above-mentioned five regression models were calculated using five different typical values computed from EMG signals during each stimulus:

1. The averaged values of EMG activity during the stimulus.
2. The peak values of EMG activity during the stimulus.
3. The baseline corrected (500 ms prestimulus baseline) average values of EMG activity during the stimulus. The average values during the baseline periods were subtracted from the average values of the corresponding stimulus periods.
4. The baseline corrected (500 ms prestimulus baseline) peak values of EMG activity during the stimulus, baseline corrected with the peak values of the baseline period. The peak values during the baseline periods were subtracted from the peak values of the corresponding stimulus periods.
5. The baseline corrected (500 ms prestimulus baseline) peak values of EMG activity during the stimulus, baseline corrected with the averaged values of the baseline period. The average values of the baseline periods were subtracted from the peak values of the corresponding stimulus periods.

2.4.2. Categorical models

There were two types of categorical models. Models that classified the estimation of subjects' emotional experiences in two categories (i.e. negative and positive) and models that classified them in three categories (i.e. negative, neutral, and positive). In all, there were 45 models. For dichotomic classification there were 30 models, and for classification in three categories there were 15 models. In both types of these estimation classification models we used the above listed frequently used EMG values for the three signals; *zygomaticus major* EMG, *corrugator supercilii* EMG, and the EMG difference score. The models were formed so that the subject's EMG responses in the calibration phase were sorted by the amplitude of the EMG response and divided into two or three similarly sized categories.

In the case of classifying into two categories, 15 models were formed from EMG responses to and ratings of positive, neutral and negative calibration stimuli and 15 models were formed from EMG responses to and ratings of positive and negative calibration stimuli only. In each model the median value was used as a limit value determining the output of the model in the testing phase. For the models based on *corrugator supercilii* activity, if the activity during a stimulus in the testing phase was greater than the median value in the calibration phase, the model estimated the subject's experience during the stimulus as a negative emotional experience. If the activity during a stimulus was smaller than the median value in the calibration phase, the model estimated the subject's experience during the stimulus as a positive emotional experience. For the models that were based on *zygomaticus major* activity or the EMG difference score, if the activity was greater than the median value in the calibration phase, the model estimated the subject's experience during the stimulus as a positive emotional experience. Similarly, if the activity during a stimulus was smaller than the median value in the calibration phase, the model estimated the subject's experience during the stimulus as a negative emotional experience.

In the case of classifying into three categories, the 1/3 and 2/3 percentile values were used as the limiting values between the three categories in the testing phase. For example, if *corrugator supercilii* activity during a stimulus in the testing phase was greater than the 2/3 percentile value in the calibration phase, the model estimated the subject's experience during the stimulus as a negative emotional experience. If the activity was smaller than the 1/3 percentile value in the calibration phase, the experience was estimated as positive, and if the activity was between the 1/3 and the 2/3 percentile values, the experience was estimated as neutral. For the models based on *zygomaticus major* activity or the EMG difference score, if the activity during a stimulus in the testing phase was greater than the 2/3 percentile value in the calibration phase, the model estimated the subject's experience during the stimulus as a positive emotional experience. If the activity was smaller than the 1/3 percentile value in the calibration phase, the experience was estimated as negative, and if the activity was between the 1/3 and the 2/3 percentile values, the experience was estimated as neutral.

2.5. Equipment

The experiment was run on a PC computer with an AMD Athlon™ XP 1800+ processor under the Windows XP Professional operating system. A 19″ Samsung

SyncMaster 959NF monitor with 1024×768 resolution was used as a display. The EMG activity was recorded with a Grass® Model15™ 8-channel differential amplifier and Link15™ 2.2 software running on a Pentium III 500 MHz PC computer under the Windows 98 operating system. The sampling rate of the system was 2000 Hz. The PolyVIEW™ PRO/32 2.0 software was used for determining the interelectrode impedances. The interelectrode impedances were <10 kΩ for all subjects. For a bipolar EMG recording, In Vivo Metric Ag/AgCl (E220X) surface electrodes filled with electrode paste were used. The EMG signal was amplified with high-pass and low-pass filters set at 100 and 1000 Hz, respectively. The signal was amplified 20,000 times for *zygomaticus major* and 50,000 times for *corrugator supercilii*. Eye blinks were monitored with electrodes attached above and below the subject's right eye. This signal was amplified 20,000 times and filtered with a high-pass filter set at 0.3 Hz and a low-pass filter set at 30 Hz.

The EMG measurement PC was connected to the subject PC with a crossover network cable, and the data was sent using the User Datagram Protocol (UDP). UDP packets were sent every 0.025 s from the measurement PC to the subject PC. The subject PC was able to process the data synchronously with the stimulus presentation software without delays.

2.6. Procedure

As the subject entered the laboratory, she/he was seated in a comfortable chair. The subject was told a cover story that during the experiment, involuntary changes in the temperature of her/his skin will be measured with surface electrodes while she/he is watching pictures and videos. A cover story was told to the subject, because the objective was to measure spontaneous facial muscle activations, and knowledge about the real purpose of the measurements might have caused the subject to exaggerate or inhibit her/his facial expressions. The subject's skin was cleaned and slightly abraded with electrode paste. Next, the electrodes were attached according to the guidelines by Fridlund and Cacioppo (1986). Electrodes were attached on the left side of the face above the *zygomaticus major* and *corrugator supercilii* muscle sites. After the electrodes had been attached, the subject was seated to view the screen so that the distance from the center of the screen to the subject's eyes was 100 cm.

The subject was told that she/he is going to see 52 pictures in the first phase of the experiment (this included both the calibration phase and the picture testing phase), and six video clips in the second phase. The subject was instructed to rate her/his affective experience after each stimulus. The ratings were given on the computer using the semantic differential method by Bradley and Lang (1994). The subject gave the ratings on a 1–9 scale. On that scale, one represented negative, five represented neutral, and nine represented positive affective experience. Before the actual stimuli, the subject saw three practice stimuli (one negative, one neutral, and one positive) in random order, and practiced giving valence ratings on the computer by rating the practice stimuli. After that, the experimenter left the room, and the subject rated 52 pictures. This phase lasted for about 20 min. All the pictures were displayed in random order within each block (i.e. the calibration block and the picture testing block) in full screen size for 6 s. Based on the EMG responses and subjective ratings for the calibration block, the system calculated

the estimation models after 24 pictures. The estimation accuracies of these models were then tested with the testing pictures and with the video stimuli.

After the subject had rated all the pictures, there was a short break before the video phase. After that, the subject rated the six videos. This phase lasted for less than 5 min. The videos were displayed in random order centered on the screen in 720×526 pixels size. There was a randomized pause of 6, 7 or 8 s before each picture and video, during which the screen was blank. The valence rating scale appeared on the screen without a delay after each stimulus.

2.7. Data analysis

The EMG data from *zygomaticus major* and *corrugator supercilii* were analyzed in real-time during the experiment as follows. First, the EMG data were rectified. Second, eye blinks were removed from the data. Specifically, data from all channels were discarded for data points, for which the electrical activity in the eye blink channel exceeded 50 μV. The stimuli, during which more than 50% of the data had to be removed due to blinking, were discarded from all further analyses. Third, the system computed continuously the estimation results for the 25 regression models and the 45 categorical models based on the EMG activities during the stimuli. The raw data including the EMG activities were also written to a file. When calculating the results of the experiment, the estimations formed by the system were compared to the actual subjective ratings of emotional experiences given by the subject.

In the regression analysis, if the valence scores predicted by the linear regression models were less than one, they were replaced by number one (a very negative experience) in the data. Likewise, if valences predicted by the regression models were greater than nine, they were replaced by number nine (a very positive experience) in the data. This procedure ensured that the system's valence estimations had the same range as in the subject's subjective ratings. In the categorical analyses, the subjects' ratings from 1 to 3 were categorized as negative experiences, ratings from 4 to 6 were categorized as neutral experiences, and ratings from 7 to 9 were categorized as positive experiences.

In the statistical analyses, one sample *t*-tests were used to determine, if the obtained estimation rates differed significantly from the chance level. For the statistical analyses of the EMG data, one-way repeated measures analyses of variance were used. Huynh-Feldt corrected degrees of freedom were used in the ANOVAs and paired samples *t*-tests were used in the pairwise comparisons. The pairwise comparisons were Bonferroni corrected when needed. For the subjective data, Friedman's rank tests were used to analyze the data, and Wilcoxon's matched-pairs signed-ranks tests were used in the pairwise comparisons. All the significance levels used in pairwise comparisons were two-tailed. For the regression-based models, the correlations between the subjects' own ratings and the ratings estimated by the machine were calculated separately for each model. A method was used in which the sample means computed from the machine estimations at each rank were correlated with the subject's own responses at that rank (i.e. nine pairs of observations across subjects). A similar method was used, for example, in Lang et al. (1993).

3. Results

3.1. Subjective ratings

The average experienced valence ratings and standard errors of the means (SEM) for the three stimulus categories and the three phases of the experiment are shown in Fig. 2. For the calibration stimuli, a Friedman's rank test showed a significant effect of stimulus category on the ratings of valence, $\chi^2_F(2) = 20.0$, $p < 0.001$. Pairwise comparisons (Wilcoxon's matched-pairs signed-ranks tests) showed that the ratings of valence were significantly more positive for the positive stimuli than the neutral stimuli, $Z = 2.8$, $p < 0.01$. The ratings of valence were also significantly more positive for the positive stimuli than the negative stimuli, $Z = 2.8$, $p < 0.01$, and significantly more positive for the neutral stimuli than the negative stimuli, $Z = 2.8$, $p < 0.01$.

For the testing pictures, a Friedman's rank test showed a significant effect of stimulus category on the ratings of valence $\chi^2_F(2) = 20.0$, $p < 0.001$. Pairwise comparisons (Wilcoxon's matched-pairs signed-ranks tests) showed that the ratings of valence were significantly more positive for the positive stimuli than the neutral stimuli, $Z = 2.8$, $p < 0.01$. The ratings of valence were also significantly more positive for the positive stimuli than the negative stimuli, $Z = 2.8$, $p < 0.01$, and significantly more positive for the neutral stimuli than the negative stimuli, $Z = 2.8$, $p < 0.01$.

For the testing video stimuli, a Friedman's rank test showed a significant effect of stimulus category on the ratings of valence, $\chi^2_F(2) = 18.0$, $p < 0.001$. Pairwise comparisons (Wilcoxon's matched-pairs signed-ranks tests) showed that the ratings of valence were significantly more positive for the positive stimuli than the neutral stimuli, $Z = 2.7$, $p < 0.01$. The ratings of valence were also significantly more positive for the positive stimuli than the negative stimuli, $Z = 2.7$, $p < 0.01$, and significantly more positive for the neutral stimuli than the negative stimuli, $Z = 2.7$, $p < 0.01$.

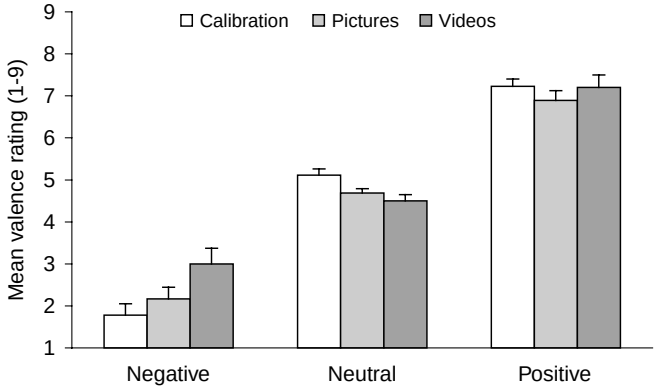

Fig. 2. The average valence ratings (and SEM) for calibration pictures, testing pictures, and testing videos.

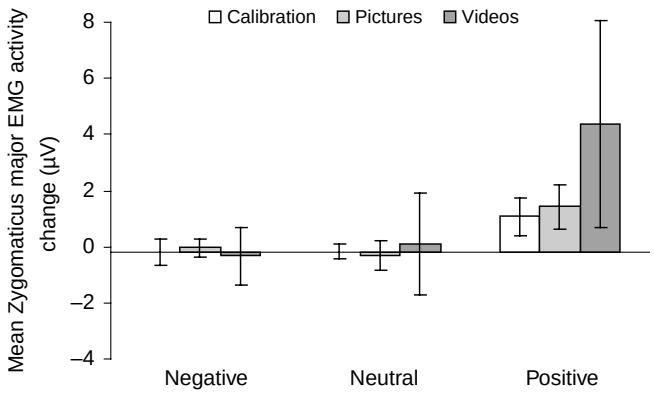

Fig. 3. Mean *zygomaticus major* EMG change from baseline (and SEM) for calibration pictures, testing pictures, and testing videos.

3.2. EMG responses

The averaged *zygomaticus major* EMG responses and are shown in Fig. 3. For the calibration stimuli, a one-way ANOVA showed a significant effect of stimulus category on *zygomaticus major* EMG, $F(1, 9)=4.1$, $p<0.05$. For the testing pictures, a one-way ANOVA also showed a significant effect of stimulus category on *zygomaticus major* EMG, $F(1, 9)=4.4$, $p<0.05$. After Bonferroni correction none of the pairwise differences became significant. For the testing videos, the effect of stimulus category on *zygomaticus major* EMG was not significant.

The averaged *corrugator supercilii* EMG responses are shown in Fig. 4. For the testing pictures, a one-way ANOVA showed a significant effect of stimulus category on *corrugator supercilii* EMG $F(1, 9)=4.4$, $p<0.05$. For the calibration stimuli and for the testing videos, the effect of stimulus category on *corrugator supercilii* EMG was not

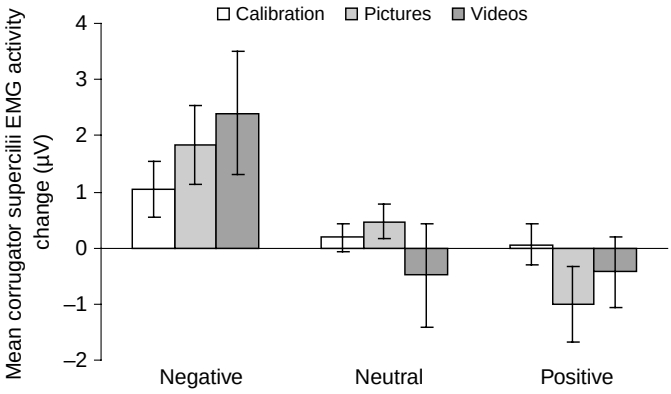

Fig. 4. Mean *corrugator supercilii* EMG activity change from baseline (and SEM) for calibration pictures, testing pictures, and testing videos.

significant. After Bonferroni correction none of the pairwise differences became significant.

3.3. Estimation results

For pictures, in classifying categorically into negative and positive responses the average estimation accuracy of all the models was 64.0% (range 36.3–70.2%). In classifying into negative, neutral, and positive responses, the average estimation accuracy for all models was 39.2% (range 34.5–46.0%). The estimations of emotional experiences using regression models correlated with an average correlation of $r=0.67$ (range 0.29–0.91) with the subjects' own ratings of experiences.

For videos, in classifying into negative and positive responses the average estimation accuracy of all the models was 62.8% (range 40.8–80.8%). In classifying into negative, neutral, and positive responses, the average estimation accuracy for all models was 40.2% (range 32.5–47.5%). Finally, the estimations of emotional experiences using regression models correlated with an average correlation of $r=0.57$ (range 0.08–0.93) with the subjects' own ratings of experiences.

Overall, the baseline corrected models and the models, which used peak values baseline corrected with peak values turned out to be the most successful. The estimation accuracies of these models are presented in Table 1. A more detailed listing of the most accurate models is presented in the Appendix A.

The differences between individual test subjects were quite large in the categorical analyses for both pictures and videos. The average estimation rates in the negative–positive analysis for pictures varied from 87.1% (range 25–100%) for the subject with the best estimation rate to 47.3% (range 22.2–66.7%) for the subject with the worst estimation rate. The average estimation rates in the negative–neutral–positive analysis for pictures varied from 56.9% (range 47.1–76.5%) for the subject with the best estimation rate to 24.8% (range 11.8–45.4%) for the subject with the worst estimation rate.

The respective individual differences for videos were as follows. The average estimation rates in the negative–positive categorical analysis for videos varied from 85.0% (range 0–100%) for the subject with the best estimation rate to 41.6% (range 25–75%) for the subject with the worst estimation rate. The average estimation rates in the negative–neutral–positive categorical analysis for videos varied from 57.8% (range 33.3–66.7%) for the subject with the best estimation rate to 25.6% (range 0–33.3%) for the subject with the worst estimation rate.

4. Discussion

The current results showed that by measuring the electromyographic activity of two facial muscles, *zygomaticus major* and *corrugator supercilii*, it was possible to estimate the subjects' affective experiences reasonably well. When using picture stimuli, most of the models used in the experiment distinguished between positive and negative affective responses at a better than chance estimation rate. When using video stimuli, the best models also had a better than chance estimation rate, and the best estimation rate between positive and negative responses was over 80%. The models that classified estimated

Table 1
Selected models and their performance in the experiment

Model	Pictures			Videos		
	Positive/negative	Positive/neutral/negative	Regression	Positive/negative	Positive/neutral/negative	Regression
	Estimation accuracy (%)	Estimation accuracy (%)	Correlation with subjects' ratings	Estimation accuracy (%)	Estimation accuracy (%)	Correlation with subjects' ratings
Zygomaticus major baseline corrected	64.6*	42.1*	0.68*	65.8	43.3	0.81**
Corrugator supercilii baseline corrected	68.6*	38.6	0.83**	80.8***	46.7**	0.87**
Difference score baseline corrected	69.1*	37.0	0.80*	72.5*	44.2	0.93**
Zygomaticus major peak values baseline corrected with peak values	64.0*	40.4**	0.49	59.2	40.8	0.70*
Corrugator supercilii peak values baseline corrected with peak values	69.5*	43.8	0.91**	74.2**	38.3	0.63
Difference score peak values baseline corrected with peak values	70.1*	46.0	0.78*	75.8**	47.5*	0.67*

The asterisks indicate the level of statistical significance: *$p<0.05$, **$p<0.01$, ***$p<0.001$.

experiences into three affective categories, positive, neutral, and negative, also performed better than chance for both pictures and videos. The estimations of emotional experiences with regression models correlated relatively well with the actual ratings of experiences given by the subjects themselves.

The analysis of the subjective ratings confirmed that the subjects experienced emotions as intended. For example, positive stimuli were subjectively rated as significantly more positive on the valence scale than negative and neutral stimuli. The analysis of physiological responses showed that subjects had increased *zygomaticus major* activity mostly in response to positive stimuli, and increased *corrugator supercilii* activity mostly in response to negative stimuli. Thus, the subjects' EMG responses were generally in line with those found in previous research (Dimberg, 1990; Greenwald et al., 1989; Lang et al., 1993, 1995; Larsen et al., 2003).

Previous research suggests that baseline corrected *corrugator supercilii* activity has a negative linear relation with experiences of affective valence. Overall, the models based on the activity of that muscle were also among the most accurate models in the current experiment. However, our results also suggest alternative models that might be used with similar or possibly even better results. While models based on *corrugator supercilii* were most successful with the dynamic video stimuli, models based on *zygomaticus major* performed even better in some analyses. In addition, the difference scores, which were calculated by subtracting the *corrugator supercilii* activity from the *zygomaticus major* activity, had relatively high estimation rates. In general, the estimation rates were highest either using a baseline correction or using difference score peak values that were baseline corrected by peak values. These findings indicate that models estimating psycho-emotional experiences on the basis of facial activity can be created in several ways and they are worth further future research.

We used affective still images and soundless video clips as stimuli in this experiment. It is very likely that with audiovisually integrated stimulus materials, for example, video clips with sound, the emotional responses would be enhanced and results in improved estimation results. Further, improvements are likely to become achievable when the context of use is considered even at a modest level. In real user environments, in which events are more meaningful to the user, such as success or failure in an important computerized task (e.g. Partala and Surakka, 2004) the estimation rates can be expected to be higher. Thus, when assessing the importance of the current results, it should be noted that in this experiment the estimation was carried out by just analyzing the facial expression data.

As noted in earlier studies (e.g. Ekman, 1985; Cohn et al., 2002) we also found that the individual differences were quite large. This suggests that the success of emotion estimation from facial expressions will depend on the particular user's facial behavior. We note further that some of our subjects showed only minor activations of the *zygomaticus major* and *corrugator supercilii* muscles. For these persons, it is possible that emotion estimation is only successful in response to very strong affective stimuli or events. On the other hand, some subjects' facial responses were quite clear using both the *zygomaticus major* and *corrugator supercilii* muscles, and the estimation worked quite well even with picture stimuli and simple analysis methods.

In addition to the differences between subjects, the ranges of accuracies for the different models were also large. This was mainly due to a few models, which performed rather poorly, while the majority of the models estimated emotions relatively well.

In this paper, we have presented methods for estimating emotional experiences based on facial activity. One drawback of the present methods is that electrodes with wires need to be attached to the subject. However, there are cases, in which the benefits of using the technology are likely to transcend this limitation. For the people who have suffered serious injuries so that they may even be unable to speak, methods for emotional communication would be a tremendous asset (e.g. Surakka et al., 2004). The current status of electrode technology is also changing rapidly. In the near future surface electrode technology can be implemented in adhesive tape and the signal can be transmitted via lightweight radio transmitters. First, prototypes of wireless electrode technology have been developed, for example, by the Wireless User Interfaces Consortium (Wireless User Interfaces Consortium, 2005).

We further note that physiological signals have been studied increasingly in the context of HCI. Well-known examples of potentially HCI related physiological signals were presented in the book by Picard (1997). Even then, the analysis and estimation of user emotions on the basis of physiological responses was recognized as an important factor for improving the quality of HCI. Since, then several applications have been developed. Existing emotion-sensing devices include, for example, the emotion mouse, which measures skin temperature, galvanic skin response, and heart rate (Ark et al., 1999) in order to detect the user's emotions. Healey (2000) developed methods for recognizing car driver's stress in traffic based on a variety of different physiological measures. In respect to facial muscle activity detection Scheirer et al. (1999) developed expression glasses, which visualized the wearer's interest and confusion levels based on the activations of two facial muscles, *corrugator supercilii* and *frontalis*. The methods presented in this paper offer a noteworthy extension to these and other currently existing methods for estimating emotion from physiological measures.

When assessing the practical implications of the current results, it has to be noted that facial expressions are the most important communication channel in human–human communication, when a person evaluates another person's affective valence (Mehrabian, 1968). Similarly, the computer could detect the user's affective facial responses during human–computer interaction, which would enable real-time affective communications with computers. Advanced methods for analyzing emotion-related activations of facial muscles could also be valuable in the evaluation of human–computer interaction. Ward and Mardsen (2003) have suggested that physiological measurements are also potentially useful in usability evaluation. In line with this we suggest that taken together the findings on the association of emotions and facial muscle activity it is very likely that measuring facial EMG, especially *corrugator supercilii* and *zygomaticus major* activity, can objectively reveal both usability problems and moments of user satisfaction. Measuring *corrugator supercilii* activity could also reveal car drivers' stress as in Healey (2000) or, for example, task-related stress during computer use.

Being able to estimate the user's emotions can offer new possibilities for enhancing the quality of interaction. There is evidence that synthetic affective feedbacks or interventions can significantly regulate the user's emotional experiences, enhance cognitive processing, and significantly regulate physiological activity. These results show that integrating emotions into HCI can improve the quality of interaction on subjective, behavioral, and physiological levels (Aula and Surakka, 2002; Partala and Surakka, 2004). By estimating the user's emotions the computer can evaluate itself if its communication has any effects on the user. By this way HCI comes still closer to the human–human communication, in which we make observations and inferences from the other interactants purely on the basis of non-verbal exchanges.

Recently, there has been some discussion on whether affective human–computer interaction should be based on voluntary emotion expressions or self-adaptive systems automatically estimating spontaneous emotions (e.g. Ward and Marsden, 2004). Our position is that methods for self-adaptive affective human–computer interaction are worth developing and studying further. Their role in future human–computer interaction can only be understood by developing and evaluating real systems that are reliable enough to enable emotion estimation with accuracy comparable or even superior to that of humans. The current work contributes towards this research goal. It should also be noted that the methods described in this paper could also be used to recognize voluntary affect expressions using facial muscles, probably with much higher accuracies than reported in the current paper for the estimation of spontaneous emotions.

In sum, the current results showed that it was possible to estimate the user's subjective affective experience based on the activations of the user's facial muscles with reasonable accuracy. Especially the categorical estimation to positive and negative responses worked rather well and could be useful in practical systems. Using our system, the estimation can be done in real time, which means that in subsequent studies the computer could also change its real time behavior in a socially meaningful way according to the estimation results. We believe that these results are a step toward the development of computers, which have human-like social and emotional capabilities, such as the ability to infer affective experiences from facial expressions.

Acknowledgements

This research was supported by The Academy of Finland, project numbers 1202183 and 177857.

Appendix A

The most successful models

The models with the highest estimation rates for pictures ($*p < 0.05$, $**p < 0.01$, the level of statistical significance)

Model	Estimation %
Difference score of peak values baseline corrected with peak values	70.1*
Difference score baseline corrected	70.0*
Difference score of peak values baseline corrected with peak values, model calculated from positive and negative calibration stimuli only	69.8*
Zygomaticus major average values	69.7**
Corrugator supercilii peak values baseline corrected with peak values	69.5*
Difference score baseline corrected	69.1*
Zygomaticus major peak values	68.9**
Corrugator supercilii baseline corrected	68.6*
Zygomaticus major, model calculated without neutral stimuli	68.6*
Zygomaticus major peak values, model calculated from positive and negative calibration stimuli only	67.8*

The models with the highest estimation rates for pictures (*$p<0.05$, **$p<0.01$, the level of statistical significance)

Model	Estimation %
Difference score of peak values baseline corrected with peak values	46.0
Corrugator supercilii peak values baseline corrected with peak values	43.8
Zygomaticus major baseline corrected	42.1*
Corrugator supercilii peak values	41.9
Zygomaticus major peak values baseline corrected with peak values	40.4**

The regression models with the highest correlations with the subjects' ratings for pictures (*$p<0.05$, **$p<0.01$, the level of statistical significance)

Model	Correlation
Corrugator supercilii peak values baseline corrected with peak values	0.91**
Corrugator supercilii peak values baseline corrected with peak values, calculated from 1–5	0.90**
Corrugator supercilii baseline corrected	0.83**
Difference score baseline corrected	0.80*
Corrugator supercilii baseline corrected, calculated from 1–5	0.80*

The models with the highest estimation rates for videos (*$p<0.05$, **$p<0.01$, ***$p<0.001$, the level of statistical significance)

Model	Estimation %
Corrugator supercilii baseline corrected	80.8***
Difference score of peak values baseline corrected with peak values	75.8**
Difference score of peak values baseline corrected with peak values, model calculated from positive and negative calibration stimuli only	75.8**
Difference score baseline corrected, model calculated from positive and negative calibration stimuli only	75.0*
Corrugator supercilii peak values baseline corrected with peak values	74.2**

(*continued on next page*)

Model	Estimation %
Difference score baseline corrected	72.5*
Zygomaticus major baseline corrected, model calculated from positive and negative calibration stimuli only	70.8
Difference score of peak values baseline corrected with average values	67.5
Zygomaticus major baseline corrected	65.8
Corrugator supercilii peak values	65.8

The models with the highest estimation rates for videos (*$p < 0.05$, **$p < 0.01$, the level of statistical significance)

Model	Estimation %
Difference score of peak values baseline corrected with peak values	47.5*
Corrugator supercilii baseline corrected	46.7**
Corrugator supercilii peak values	45.0
Corrugator supercilii peak values baseline corrected with average values	45.0
Difference score baseline corrected	44.2

The regression models with the highest correlations with the subjects' ratings for videos (**$p < 0.01$, the level of statistical significance)

Model	Correlation
Difference score baseline corrected	0.93**
Corrugator supercilii baseline corrected	0.87**
Corrugator supercilii peak values baseline corrected with average values	0.82**
Corrugator supercilii peak values	0.82**
Zygomaticus major baseline corrected	0.81**

References

Ark, W., Dryer, D., Lu, D., 1999. The Emotion Mouse. In: Bullinger, H.J., Ziegler, J. (Eds.), Human–Computer Interaction: Ergonomics and User Interfaces. Lawrence Erlbaum, London, pp. 818–823.

Aula, A., Surakka, V., 2002. Auditory Emotional Feedback Facilitates Human–Computer Interaction, Proceedings of HCI 2002. Springer, Berlin pp. 337–349.

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

Bradley, M.M., Lang, P.J., 2000. Affective reactions to acoustic stimuli. Psychophysiology 37, 204–215.

Cohen, I., Sebe, N., Chen, L., Garg, A., Huang, T.S., 2003. Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding 91 (1–2), 160–187.

Cohn, J.F., Schmidt, K., Gross, R., Ekman, P., 2002. Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In: Proceedings of the International Conference on Multimodal User Interfaces 2002.

Dailey, M.N., Cottrell, G.W., Padgett, C., Adolphs, R., 2002. EMPATH: a neural network that categorizes facial expressions. Journal of Cognitive Neuroscience 14, 1158–1173.

Dimberg, U., 1990. Facial electromyography and emotional reactions. Psychophysiology 19, 643–647.

Ekman, P., 1985. Telling Lies. W.W. Norton, New York.

Ekman, P., 1993. An argument for basic emotions. Cognition and Emotion 6 (3), 169–200.

Essa, I.A., Pentland, A.P., 1997. Coding, analysis, interpretation and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 757–763.

Fridlund, A.J., Cacioppo, J.T., 1986. Guidelines for human electromyographic research. Psychophysiology 23 (5) , 567–589.

Greenwald, M.K., Cook III., E.W., Lang, P.J., 1989. Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. Journal of Psychophysiology 3, 51–64.

Healey, J., 2000. Wearable and Automotive Systems for the Recognition of Affect from Physiology. MIT PhD Thesis—Electrical Engineering and Computer Science Department.

Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O., 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology 30, 261–273.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1995. International Affective Picture System (IAPS): Technical Manual and Affective Ratings. The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Larsen, J.T., Norris, C.J., Cacioppo, J.T., 2003. Effects of positive and negative affect on electromyographic activity over *zygomaticus major* and *corrugator supercilii*. Psychophysiology 40, 776–785.

Lisetti, C.L., Schiano, D.J., 2000. Automatic facial expression interpretation: where human–computer interaction, artificial intelligence and cognitive sciences intersect. Pragmatics and Cognition 8 (1), 185–235.

Mehrabian, A., 1968. Communication without words. Psychology Today 2 (9), 52–55.

Oliver, N., Pentland, A., Berard, F., 2000. LAFTER: a real-time face and lips tracker with facial expression recognition. Pattern Recognition 33, 1369–1382.

Partala, T., Surakka, V., 2003. Pupil size as an indication of affective processing. International Journal of Human–Computer Studies 59, 185–198.

Partala, T., Surakka, V., 2004. The effects of affective interventions in human–computer interaction. Interacting with Computers 16 (2), 295–309.

Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge, MA.

Scheirer, J., Fernandez, R., Picard, R.W., 1999. Expression Glasses: A Wearable Device for Facial Expression Recognition, Proceedings of CHI'99, Extended Abstracts 1999 pp. 262–263.

Smith, E., Bartlett, M.S., Movellan, J., 2001. Computer Recognition of Facial Actions: A Study of Co-articulation Effects, Proceedings of the Eigth Annual Joint Symposium on Neural Computation 2001.

Surakka, V., Hietanen, J.K., 1998. Facial and emotional reactions to Duchenne and non-Duchenne smiles. International Journal of Psychophysiology 29, 23–33.

Surakka, V., Illi, M., Isokoski, P., 2004. Gazing and frowning as a new technique for human–computer interaction. ACM Transactions on Applied Perception 1 (1), 40–56.

Ward, R.D., Mardsen, P.H., 2003. Physiological responses to different WEB page designs. International Journal of Human–Computer Studies 59 (1–2), 199–212.

Ward, R.D., Marsden, P.H., 2004. Affective computing: problems, reactions and intentions. Interacting with Computers 16 (4), 707–713.

Wireless user interfaces consortium, 2005. http://www.cs.uta.fi/hci/wtpc/results.html (last accessed 05/05/05).