Continuous estimation of emotional change using multimodal affective responses

Kenta Masui
Chiba University
Chiba, Japan
k_masui@chiba-u.jp

Takumi Nagasawa
Chiba University
Chiba, Japan
takumi-n9stillseigoro-future@chiba-u.jp

Hirokazu Doi
Kokushikan University
Tokyo, Japan
hdoi@kokushikan.ac.jp

Norimichi Tsumura
Chiba University
Chiba, Japan
tsumura@faculty.chiba-u.jp

Abstract

Emotions have a significant effect on our daily behavior, such as perception, memory, and decision making. For this reason, interest in considering the emotions of the user in a human-computer interface has recently increased. This is important for future interface applications, which are expected to operate in harmony with humans. In this paper, we present our approach to instantaneously detecting the emotions of video viewers from remote measurement using an RGB camera. Facial expression and physiological responses, such as heart rate and pupil diameter, were measured by analyzing facial videos. We also verified the effectiveness of the contactless measurement by acquiring electroencephalogram signals using a contact-type electroencephalograph. By combining the measured responses into multimodal features and using machine learning, we showed that the results of emotion estimation were better than estimates made from only single-mode features.

1. Introduction

Estimation of emotions is one of the major concerns in the development of human-computer interfaces for robots. The demand for customer service robots and welfare support robots is increasing in Japanese society, where population decline is regarded as a problem. These service robots are required to interact naturally with humans. In this situation, if the robot is able to be mindful of the customer and behave in a manner that considers the customer's state of mind, the robot may offer better service than humans. Extensive research has been conducted for this purpose, and it has focused particularly on expressions such as the user's gaze direction [1], head position, facial expression (FE) [2], behavior, and speech [3]. However, the limitation of these approaches is that they cannot recognize emotions accurately when FE and actions are intentionally falsified. For example, people can express a fake smile. For this reason, research that combines physiological information, such as electroencephalography (EEG) and heart rate (HR), has been conducted to estimate emotions. Physiological psychology has shown that there is a strong correlation between human emotions and the physiological response of the autonomic nervous system. In addition, physiological responses cannot be controlled intentionally. Measuring physiological responses from the human face and body using a camera is a new research area that has grown rapidly in recent years, and it provides a natural way for robots to use their cameras to estimate emotions without physical contact.

Most research on automatic emotion recognition has focused on the analysis of six individual basic emotions [4] (happiness, sadness, surprise, fear, anger, and disgust). However, these basic emotions are independent and may not be able to explain the complexity of affective conditions very well. For this reason, research has been conducted to measure human affects such as excitement [5], stress [6], concentration [7], and relaxation [8]. In contrast, detection of continuous or dimensional emotional expressions is based on the assumption that emotions can be described in a continuum without being divided into distinct groups [9, 10]. Because each emotion can be expressed continuously, measuring continuous changes in emotions is useful. To achieve a natural and intuitive interaction between humans and robots, it is essential that time-continuous emotion prediction analyze subtle and complex affective states of humans over time.

The goal of this study was to continuously detect emotions from FEs and the physiological responses of pupil

diameter and HR. In addition, the effectiveness of contactless measurement was verified by comparing its accuracy with EEG signals obtained using a contact-type electroencephalograph. First, we collected facial videos and EEG data for subjects watching a set of emotion-evoking videos. During the viewing, subjects continuously recorded their feelings while watching the videos. The continuous annotations served as the ground truth for our continuous emotion detection system. Next, FEs and physiological responses (HR and pupil diameter) were remotely measured by analyzing the facial videos. Finally, by combining the measured responses as multimodal features and using an extreme learning machine (ELM), we performed continuous emotion detection. We evaluated the emotion detection results with a 10-fold cross-validation strategy using an average correlation coefficient and the concordance congruence coefficient (CCC). To the best of our knowledge, this is the first attempt to detect continuous emotions, in both time and dimension, using subjective assessment annotations as the ground truth label.

The rest of this paper is structured as follows. We describe related work with multimodal emotion recognition and temporal dynamics of emotions in Section 2. The experimental environment is outlined in Section 3. Contactless measurement features obtained from the camera and EEG features are given in Section 4. Section 5 describes the ELM method for emotion recognition. The experimental results are presented and discussed in Section 6. Finally, the paper is concluded in Section 7.

2. Background

A major attempt to advance the state of the art in continuous emotion detection was the Audio/Visual Emotion Challenge (AVEC) 2012 [11]. The goal of the AVEC 2012 challenge was to detect the continuous dimensional emotions using audiovisual signals. Since then, many researchers have tried to detect continuous dimensional emotions using various modalities. Mohammad *et al*. used FE and brain waves [12]. Bugnon *et al*. used pulse waves measured by a contact-type sensor [13]. A comprehensive review of continuous emotion detection is given in [14]. The ground truth label in these studies was a weighted average of objective evaluations by multiple annotators.

Meanwhile, McDuff *et al*. measured the level of smiling during viewing of a video advertisement to assess the subjects' preference for the content [15]. Chakraborty *et al*. detected viewer interest automatically using FE and HR [16]. The ground truth label in these studies was determined from subjective annotations and ratings. Because we are interested in detecting the emotions we feel, we focus on emotions determined from subject annotations and ratings.

Also, the performance of multimodal emotion recognition is generally better than a single modality. Many studies have shown the effectiveness of combining FE responses and biological responses [16, 17, 18]. These results demonstrated that HR signals could provide complementary information during FE detection.

3. Data Set and Annotations

3.1. Participants

Subjects were recruited on campus from 35 healthy right-handed students, comprising 23 men and 12 women with an age range of 21 to 25 years. All subjects underwent a physical evaluation to screen out chronic diseases and mental disorders. These were assessed by the Japanese versions of the autism-spectrum quotient [19], the Toronto alexithymia scale [20], the Beck depression inventory [21], and the State-Trait anxiety inventory [22]. No subjects were excluded from this study due to mental disorder. Participants granted their written informed consent to capture facial videos and perform EEG, and all were informed of their right to discontinue participation at any time. Each subject received 2,000 Japanese yen for participation. The study procedures were approved by the Engineering Research Ethical Committee of the Chiba University, under reference number 31-08.

3.2. Stimuli Video Clips

Our study used movie clips collected from FilmStim [23] as emotional stimuli. FilmStim is a database of brief video clips intended to elicit emotional states in experimental psychology experiments. Six movie scenes were selected to cover the whole spectrum of emotions from famous commercial movies, specifically "There is Something about Mary," "American History X," "The Silence of the Lambs," "The Blair Witch Project," "A Perfect World," and "The Dead Poets Society." Each film scene is intended to elicit amusement, anger, disgust, fear, sadness, or tenderness. The duration of each movie clip is 83-279 seconds, which is long enough to evoke emotions in the viewer [24].

3.3. Data Collection

The experimental setup is shown in Figure 1. Subjects were asked to sit on a chair and place their head on a chin rest. We asked subjects to keep their body as motionless as possible during the experiment. However, facial expressions are not limited and are expressed naturally. The facial videos and EEG were recorded from the subjects watching video clips. Video FE data were recorded using a DFK33UX174 RGB camera from Argo Inc. This camera

has a resolution of 1,024×768 pixels at 30 frames per second. Video clips were played randomly for each subject, and 210 videos were recorded in total. EEG signals were acquired from 16 active electrodes on an international 10-20 system using an OpenBCI Ultracortex Mark IV (electrode impedance < 10 kΩ, 1–50 Hz, and 125 samples/sec).

Subjects self-reported dynamic emotions that played out over time using software based on the dual axis rating and media annotation tool (DARMA) [15] and a joystick. DARMA is a modernized continuous measurement system that synchronizes media playback and the continuous recording of two-dimensional measurements [15]. The subjects plotted their emotional state using a joystick whose positions corresponded to the Russell's dimensional emotional expressions [9] shown in Figure 2. The level of arousal and valence elicited by watching the video clips were recorded in the range [-1,000, 1,000] at a sampling rate of 2 Hz as ground truth.

To synchronize the acquired signal with the video clip, we centrally monitored the timing of all sensors with modified DARMA and an Arduino interface. Trigger signals were sent from DARMA to the interface when the video clip was started and ended. Thus, the EEG data captured a trigger signal from Arduino. In addition, an LED light attached to the chin rest was turned on by the trigger signal. The facial videos ware automatically cropped when the reflected LED light was turned on. This allowed us to synchronize the ground truth data with all other modalities.

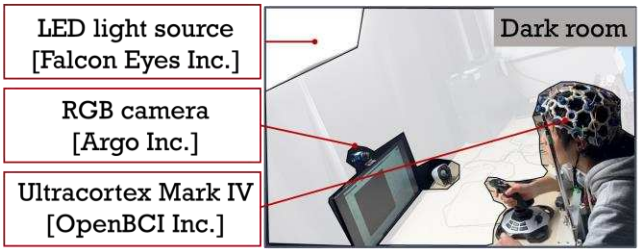

Figure 1. Experimental setup. The distance between the LED light source and subject is approximately 1.2 m. The distance between the RGB camera and subject is approximately 1.0 m.

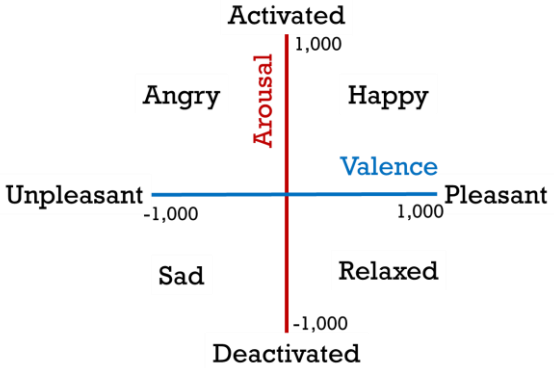

Figure 2. Russell's dimensional emotional expressions.

4. Feature Extraction

The features of FE, pupil diameter, and HR were calculated by analyzing facial videos without physical contact. The features of each channel were used individually or in combination to estimate the subject's emotions. The features used in this study and described in this section are summarized in Table 1.

Channel	Features
FE	Action unit (AU), emotion expressed by AUs
Pupil diameter	Pupil diameter ratio
HR	MeanHR, sdHR, RMSSD, pNN50, HRVti, SD1, SD2, aLF, aHF, pLF, pHF, nLF, nHF, LF/HF
EEG	Theta, alpha, beta, gamma, arousal, valence

Table 1. Features used in the estimation of the emotions (see the text for explanation of abbreviations).

4.1. Facial Expression (FE) Features

We used the OpenFace library [26] to detect FE features. First, we extracted 68 face parts that were detected from each frame of the recorded facial video. We performed a similarity transform to align all images to a common reference frame using tracked facial landmarks, with a resolution of 112×112 pixels. Also, we detected oriented edges to use a histogram of oriented gradients descriptor (HOG). Each facial image was divided into 11×11 cells, and a histogram was calculated for each pixel within each cell according to gradient strength weights. HOG is the concatenation of these histograms. Using the coordinates of the 68 facial parts and the HOG, the 17 action units (AUs) shown in Table 2 were detected by a linear kernel with support vector regression. Each AU represents a different facial muscle movement. Certain combined movements of these facial muscles pertain to an expressed emotion. For example, happiness is calculated from the combination of AU6 (cheek rise) and AU12 (lip corner pull). A complete list of these combinations and the emotion obtained is given in Table 3. Action units and the emotions they express were calculated from each frame of the facial videos.

AU	Description
1	Inner brow rise
2	Outer brow rise
4	Brow lowering
5	Upper lid rise
6	Cheek rise
7	Lid tightening
9	Nose wrinkling
10	Upper lip rise
12	Lip corner pull
14	Dimpling
15	Lip corner depression
17	Chin rise
20	Lip stretch
23	Lip tightening
25	Lip parting
26	Jaw drop
45	Blink

Table 2. Action units (AUs).

Emotion	Action Units
Happiness/joy	6 + 12
Sadness	1 + 4 + 15
Surprise	1 + 2 + 5 + 26
Fear	1 + 2 + 4 + 5 + 7 + 20 + 26
Anger	4 + 5 + 7 + 23
Disgust	9 + 15 + 25
Contempt	12 + 14

Table 3. Emotions and action units.

4.2. Pupil Features

We calculated the pupil diameter ratio by improving Filipe *et al.*'s method [27]. Figure 3 is a flowchart of the pupil diameter ratio calculation. Eye features were detected using the Haar Cascade method [28] from the captured facial video. The iris and pupil were segmented by applying a trained U-Net [29] model to the detected eye images. Next, the areas of the iris and pupil were calculated for each frame by performing contour detection and circle fitting on the output image segmented into regions. Then, the ratio between the iris and pupil areas was used as an index of pupil enlargement and contraction to avoid the problem of distance between the eyes and camera. Missing values due to blinking were interpolated linearly. The interpolated signal was resampled at 2 Hz, and a smoothing process was applied to remove noise. Finally, because the pupils of both eyes are almost the same size and respond to light in the same way, the average of the ratio of the pupil area to the iris area of both eyes was obtained.

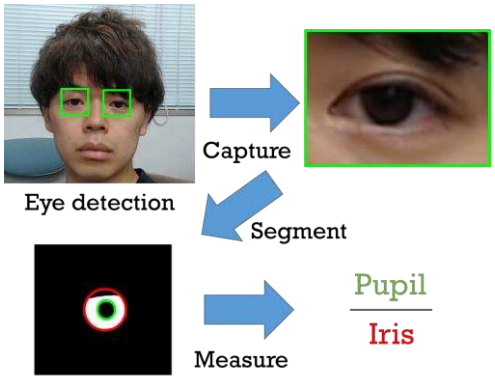

Figure 3. Flow of pupil diameter ratio calculation

4.3. Heart Rate Features

Heart rate was measured remotely using an original method based on Fukunishi *et al.*'s method [30]. Figure 4 is a flowchart for the remote HR measurement. First, from the coordinates of the detected face parts, the skin region of the face (the face region from which the eyes, nose, and mouth are removed) is taken as the region of interest. Based on the assumption that the skin is composed of two layers, melanin and hemoglobin, spatial principal component analysis was applied to extract the hemoglobin image, which is sensitive to blood volume [31]. After extraction of hemoglobin images from RGB images, the pulse wave was produced by spatially averaging the hemoglobin components in the region of interest. The detailed approach is described in [30, 31].

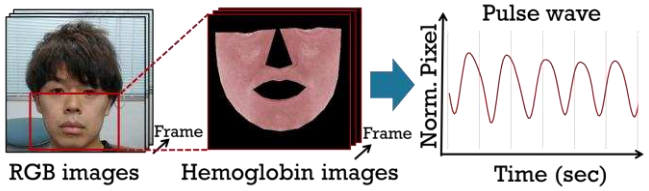

Figure 4. Remote heart rate measurement.

Next, detrending and a bandpass filter were applied to all extracted pulse waves to reduce noise. Detrending based on the smoothness prior approach [32] was applied to the waveform to eliminate low-frequency (LF) fluctuations. After applying the detrending, a band-pass filter was applied to extract the heartbeat components with a frequency between 0.75 Hz (45 bpm) and 3 Hz (180 bpm) corresponding the HR range in adults. After the band-pass filtering, the estimation of pulse rate variability (PRV) was refined by applying cubic spline interpolation to the filtered signal for upsampling from 30 Hz to 500 Hz. HR features were obtained at a 2-Hz sampling rate with a moving window of 30 sec and 0.5-sec increments. This window

length makes possible to have enough data for feature extraction without losing time resolution [33].

The interbeat intervals (IBIs) were calculated by detecting the peaks in the moving window. The time-domain method can be easily executed because it directly analyzes the IBI. The average of the HR (meanHR) and standard deviation of the HR (sdHR) were the most easily obtainable indexes. They were calculated by dividing 60 by the average IBI. The root mean square of the successive difference (RMSSD) reflects the short-term variation. Furthermore, pNN50, which is the number of IBIs for which the successive difference is 50 msec or longer relative to the total number of consecutive IBIs, was also used as an indicator of parasympathetic nerve activity.

In addition to these statistical features, we used the geometric-domain method, which is based on calculations taken from a geometric pattern whose basis lies within the IBI series. The most common geometric pattern used is the IBI histogram. The HR variability triangular index (HRVti) is a value obtained by dividing the area integral value (total number of IBIs) of the histogram of the IBI by the maximum value of the histogram.

The Poincaré plot, named after Henry Poincaré (also called a first-return map), is a type of nonlinear-domain method used to quantify self-similarity [34]. Often an ellipse is fitted to the plotted data with the long axis along the line of identity defined by $y = x$. Standard deviations along the line of identity (SD2) and perpendicular to the line of identity (SD1) represent the magnitude of the major and minor axes of the ellipse, respectively. SD1 represents the standard deviation of the instantaneous beat-to-beat (short-term) variability. SD2 represents the standard deviation of continuous (long-term) variability.

In the frequency-domain method, IBI power spectral density (PSD) is analyzed. The features obtained from the PSD are commonly used as an indicator of autonomic nervous system activity. The PSD can be estimated using many methods, but methods based on fast Fourier transform (FFT) and autoregressive (AR) modeling are perhaps the most popular in spectral analysis of PRV [35]. However, both FFT and AR-based PSD estimates have prerequisites that are seldom if ever met by biological signals, such as cardiac IBI series [36]. Consequently, other methods, such as the Lomb–Scargle periodogram and methods based on wavelet transforms, are becoming popular [37, 38]. In this study, the PSD of the heartbeat RR interval was calculated using five PSD estimation methods, including Welch's method [39]. The high-frequency (HF) band (0.15–0.4 Hz) component of the PRV reflects the respiratory sinus arrhythmia affected by respiratory and parasympathetic activity. Meanwhile, the LF band (0.04–0.15 Hz) component represents the Mayer wave originating from both sympathetic and parasympathetic activity. Finally, the integral value of HF (aHF) and LF (aLF) in the PSD, the percentage of HF (pHF) and LF (pLF) in the entire PSD, the normalized values using only HF (nHF) and LF (nLF), and the ratio of LF to HF (LF/HF) were used as features.

4.4. EEG Features

Figure 5 shows the EEG electrode distribution of the international 10-20 system. EEG signals were analyzed with the EEGLAB toolbox [40]. First, EEG signals were digitally band passed to 1–40 Hz to eliminate noise from the power source. Next, EEG artifacts were corrected by Winkler *et al.*'s method [41] using independent component analysis. Thereafter, channels with excessive artifacts were interpolated. Remaining artifacts were removed manually.

It is known that the PSD of EEG signals in various bands is correlated with emotion. An FFT algorithm was applied to all extracted artifact-free EEG signals. The PSD was obtained at a 2-Hz sampling rate with a moving window of 30 sec and a step of 0.5 sec. The logarithms of the PSD from the theta (4 Hz < f < 8 Hz), alpha (8 Hz < f < 12 Hz), beta (12 Hz < f < 30 Hz), and gamma (30–40 Hz) bands were extracted to serve as features. In addition, alpha waves in the prefrontal cortex are more dominant in the relaxed state and alpha activity was associated with brain inactivity. Thus, the alpha band is a reasonable indicator of the arousal state of a person. Concretely, the arousal level was computed as follows:

$$Arousal = \ln(\alpha P4) + \ln(\alpha P3). \quad (1)$$

The asymmetry of the two cortical hemispheres was used to determine the valence level. Davidson has demonstrated that the left frontal area is associated with more positive effects and memories, and the right hemisphere is more involved in negative emotions [42]. The F3 and F4 positions are used most often for looking at this alpha activity related to valence, as they are located in the prefrontal lobe, which plays a crucial role in emotion regulation and conscious experience. Valence values were computed by comparing the alpha power α in channels F3 and F4. Concretely, the valence level was computed as

$$Valence = \ln(\alpha F4) - \ln(\alpha F3). \quad (2)$$

Arousal and valence computation were adapted from Ramirez *et al.*'s method [43], where the authors show that the computed arousal and valence values indeed contain meaningful emotional information. In total, the number of EEG features of a trial with 16 electrodes and 4 bands is 16 × 4 + 2 = 66 features.

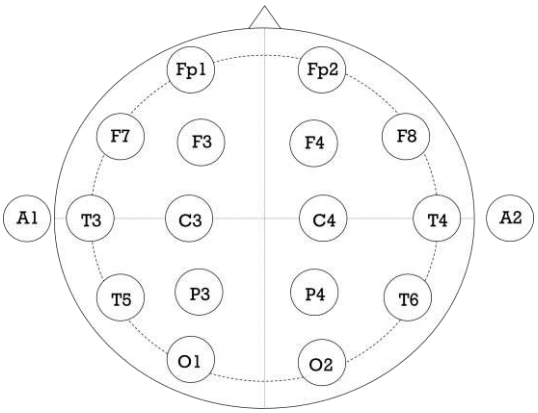

Figure 5. Electrode distribution.

5. Dimensional Affect Prediction

5.1. Extreme learning machines

Because the ELM algorithm [44] can be easily implemented, it tends to have the smallest training error, obtain the smallest norm of weights with good generalization performance, and run extremely fast. These advantages differentiate it from other popular single hidden layer feedforward neural networks. Further, it tends to have good generalization performance for feedforward neural networks. Given the input data $X = [x_1, x_2, \dots, x_i]$ and target data $T = [t_1, t_2, \dots, t_i]$, the hidden layer is represented as follows:

$$h_j = \phi\left(v_j^T x + b_j\right), \tag{3}$$

where ϕ is the activation function, v_j is the random weight between the input layer and the hidden layer generated by random numbers, and b_j is the bias. The hidden layer is given as $H = [h_1, h_2, \dots, h_N]$. When the weight between the hidden layer and the output layer is W, the output Y is given as follows:

$$Y = HW. \tag{4}$$

To find W, the problem can be stated as follows:

$$\underset{W}{\text{minimize}} \|HW - T\|_2. \tag{5}$$

This is a least-squares optimization problem.

5.2. Evaluation Metrics

Finding optimal evaluation metrics for dimensional and continuous emotion prediction and recognition remains an open research issue [45]. In this study, the accuracy was evaluated using Pearson's correlation coefficient (COR) and the CCC [46].

COR is an indicator of the strength of the linear relationship between prediction and ground truth. Let $\hat{y}$ be the prediction and y be the ground truth. COR is defined as follows:

$$COR = \frac{COV_{y\hat{y}}}{\sigma_y \sigma_{\hat{y}}}, \tag{6}$$

where σ is the standard deviation and COV is the covariance.

The CCC is a statistical measure of the agreement between the values of two equally sized vectors $\hat{y}$ and y. It combines COR with the squared difference. The CCC is defined as follows:

$$CCC = \frac{2p\sigma_y \sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{7}$$

where p is the COR between the two vectors, σ^2 is the variance of the respective vector, and μ is its mean value. The CCC can have values between -1 and 1, where 1 means a strong similarity and -1 means dissimilarity. Unlike the COR, the CCC penalizes predictions that are well correlated with the ground truth but shifted in value in proportion to their deviation. This property makes the CCC metric meaningful for the evaluation of our two-dimensional emotion labels, that is, $\in [-1,1]^2$. The correlation between the predicted emotion and the true emotion is considered along with the prediction value's divergence from the real value.

6. Experimental Results and Discussion

Continuous detection of emotions is performed by an ELM using the calculated features and annotations recorded continuously. These annotations were normalized to [-1 1] for use as the ground truth. In addition, feature-level fusion was used to fuse the functions of multimodal processing. To create a generalizable model, learning and regression were carried out by K-fold cross validation. In the K-fold cross validation, the sample group is divided into K parts. One of them is a test case and the remaining K – 1 parts are training cases. In cross validation, each sample group is verified K times as a test case. The average of the K evaluations is obtained in this way, and a single estimate is calculated. During the validation, parameter K was 10.

The COR results of continuous emotion recognition are shown in Figure 6. The CCC results are shown in Figure 7. An example of continuous detection of valence using FE, pupil diameter, and HR is given in Figure 8. Based on the results in Figures 6 and 7, we can make three points.

First, we discuss the overall estimation result of arousal and valence. As a general tendency, the degree of valence was more accurately estimated than the degree of arousal. This suggests that the ground truth labels due to subjective dependence affect the estimation accuracy. Because the subjects carefully and quietly watch the movie clips during the experiment, it is difficult for the subject to rate arousal. Nevertheless, it is presumed that the evaluation of valence may be intuitive and easy to perform.

Second, we discuss the estimation results of each modal. The fusion of contactless measurement modalities achieved better results than those of methods using conventional EEG and FE separately. This result suggests that the multimodal approach is effective. In addition, when comparing single physiological signal features, EEG achieved better arousal measurement results than did HR or pupil diameter. However, the HR and pupil diameter measurement achieved better results than did EEG in terms of valence. This is consistent with Ikeda *et al.*'s finding that EEG is suitable for estimating human cognitive arousal, while HR is suitable for estimating the autonomic nervous state (pleasant or unpleasant) [47]. But we also have to pay attention to the credibility of the EEG signal. Mohammad *et al.* [12] point out that EEG signals can be affected by the contamination of facial muscle activity. We will investigate this aspect further in the future.

Third, we discuss the comparison with other work. A direct comparison of the performance to the other works [11, 12, 13, 14] is not possible due to the nature of the database and the experimental environment. In particular, the methods for attaching ground truth labels differs greatly in terms of subjective evaluation by subjects and objective evaluation by annotators. In this study, we were affected by the problem of subjectivity. However, as shown in Figure 6, we were able to capture the tendency of emotional change in some cases, so it may be possible to improve accuracy by collecting more data to reduce the effects of individual differences.

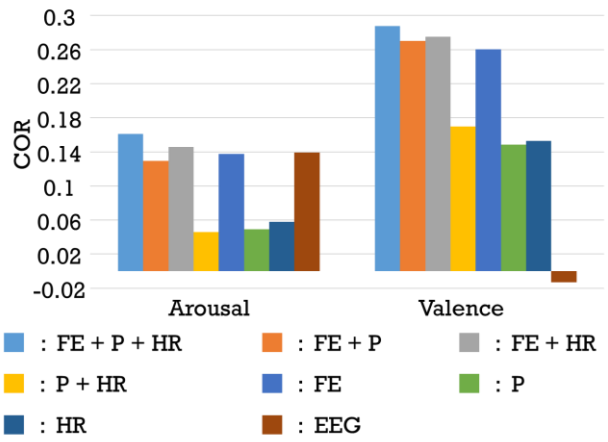

Figure 6. COR results of continuous emotion recognition. FE = facial expression, P = pupil, and HR = heart rate.

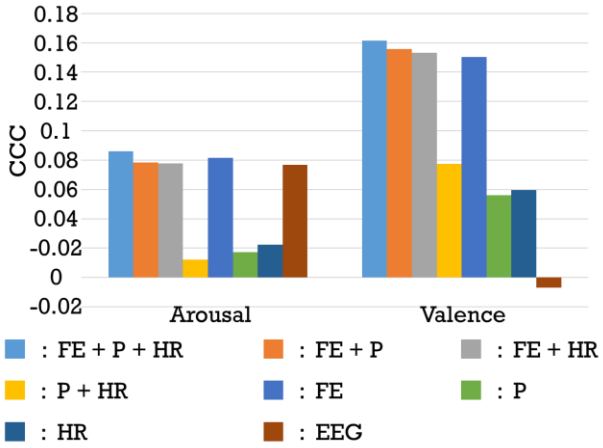

Figure 7. CCC results of continuous emotion recognition. Abbreviations are the same as in Figure 6.

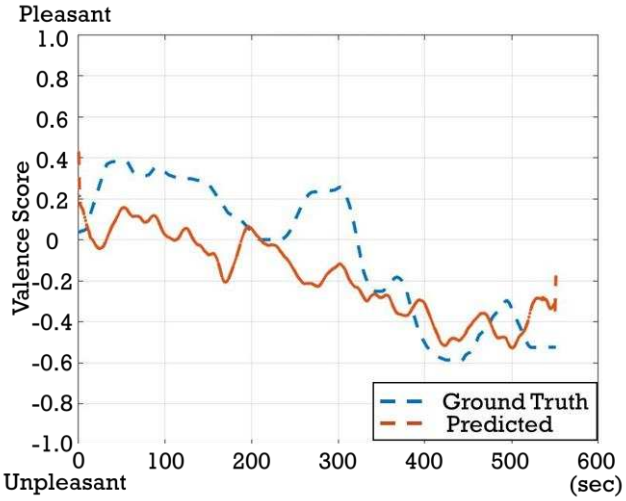

Figure 8. Example of continuous detection of valence using facial expression, pupil diameter, and heart rate. In this case, increased discomfort due to movie stimulus is tracked.

7. Conclusion and Future Work

We proposed a method to continuously detect emotions from FE and physiological responses. Our contributions are discussed in this section. First, measurement with contactless features was more accurate than measurement with EEG features, indicating the effectiveness of contactless measurement. It is more realistic to use a camera for engineering applications because EEG is sensitive to noise. Next, compared with measurement using only FE, the combination of multiple physiological signals provided a more accurate estimation, indicating the effectiveness of multimodal analysis.

Future tasks are to further improve the method's accuracy. Analyzing a large number of subjects using crowdsourcing can be expected to help reduce the influence of self-assessment bias [15, 17]. Alternatively, we will

consider using objective measures of emotions with brain waves as ground truth [48]. The continuously recorded observations should be standardized to improve the ground truth.

References

[1] Yukiko I. Nakano and Ryo Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. Proceedings of the 15th International Conference on Intelligent User Interfaces, 139–148, 2010.

[2] Won-Kyung Song *et al*. Visual servoing for a user's mouth with effective intention reading in a wheelchair-based robotic arm. Proceedings 2001 IEEE International Conference on Robotics and Automation, 4:3662–3667, 2001.

[3] Hanan Salam and Mohamed Chetouani. Engagement detection based on multi-party cues for human robot interaction. 2015 International Conference on Affective Computing and Intelligent Interaction, 341–347, 2015.

[4] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

[5] Michiko Ohkura *et al*. Measurement of "wakuwaku" feeling generated by interactive systems using biological signals. Proceedings Kansei Engineering and Emotion Research International Conference, 2293–2301, 2010.

[6] Ryota Mitsuhashi *et al*. Video-Based Stress Level Measurement Using Imaging Photoplethysmography. 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 90–95, 2019.

[7] Hamed Monkaresi *et al*. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2016.

[8] Kimberly Chu and Chui Yin Wong. Player's attention and meditation level of input devices on mobile gaming. 2014 3rd International Conference on User Science and Engineering (i-USEr), 13–17, 2014.

[9] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[10] Daniel Västfjäll *et al*. The measurement of core affect: A Swedish self-report measure derived from the affect circumplex. *Scandinavian Journal of Psychology*, 43(1):19–31, 2002.

[11] Björn Schuller *et al*. AVEC 2012: the continuous audio/visual emotion challenge. Proceedings of the 14th ACM International Conference on Multimodal Interaction, 449–456, 2012.

[12] Mohammad Soleymani *et al*. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2015.

[13] Leandro Ariel Bugnon, Rafael A. Calvo, and Diego Humberto Milone. Dimensional affect recognition from HRV: An approach based on supervised SOM and ELM. *IEEE Transactions on Affective Computing*, 2017.

[14] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

[15] Daniel McDuff *et al*. Predicting online media effectiveness based on smile responses gathered over the internet. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1–7, 2013.

[16] Prithwi Raj Chakraborty *et al*. Towards generic modelling of viewer interest using facial expression and heart rate features. *IEEE Access*, 6: 62490–62502, 2018.

[17] Kenta Masui, Genki Okada, and Norimichi Tsumura. Measurement of advertisement effect based on multimodal emotional responses considering personality. *ITE Transactions on Media Technology and Applications*, 8(1):49–59, 2020.

[18] Phuong Pham and Jingtao Wang. Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. Proceedings of the 22nd International Conference on Intelligent User Interfaces, 67–78, 2017.

[19] Akio Wakabayashi *et al*. The autism-spectrum quotient (AQ) in Japan: a cross-cultural comparison." *Journal of Autism and Developmental Disorders*, 36(2):263–270, 2006.

[20] Isao Fukunishi *et al*. Is alexithymia a culture-bound construct? Validity and reliability of the Japanese versions of the 20-item Toronto Alexithymia Scale and modified Beth Israel Hospital Psychosomatic Questionnaire. *Psychological Reports*, 80(3):787–799, 1997.

[21] Masayo Kojima *et al*. Cross-cultural validation of the Beck Depression Inventory-II in Japan. *Psychiatry Research*, 110(3):291–299, 2002.

[22] Katsuharu Nakazato and Yoshiko Shimonaka. The Japanese state-trait anxiety inventory: age and sex differences. *Perceptual and Motor Skills*, 69(2):611–617, 1989.

[23] Alexandre Schaefer *et al*. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010.

[24] Yoann Baveye *et al*. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[25] Jeffrey M. Girard and Aidan G.C. Wright. DARMA: Software for dual axis rating and media annotation. *Behavior Research Methods*, 50(3):902–909, 2018.

[26] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 1–10, 2016.

[27] Conceicao Filipe *et al*. Pupal-deep-learning. https://github.com/pupal-deep-learning/PuPal-Beta. Accessed 24 Feb 2020.

[28] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 2001.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical image Computing and Computer-assisted Intervention, 234–241, 2015.

[30] Munenori Fukunishi *et al*. Non-contact video-based estimation of heart rate variability spectrogram from hemoglobin composition. *Artificial Life and Robotics*, 22(4):457–463, 2017.

[31] Norimichi Tsumura *et al*. Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. ACM SIGGRAPH 2003 Papers, 770–779, 2003.
[32] Mika P. Tarvainen, Perttu O. Ranta-Aho, and Pasi A. Karjalainen. An advanced detrending method with application to HRV analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002.
[33] Björn W. Schuller. Acquisition of affect. *Emotions and Personality in Personalized Services*, 57–80, 2016.
[34] P.W. Kamen and Andrew M. Tonkin. Application of the Poincaré plot to heart rate variability: A new measure of functional status in heart failure. *Australian and New Zealand Journal of Medicine*, 25(1):18–26 , 1995.
[35] Gari D. Clifford *et al*. Advanced methods and tools for ECG data analysis. Boston:Artech House, 2006.
[36] John G. Proakis. Digital signal processing: principles algorithms and applications. India:Pearson Education, 2001.
[37] C. Lévy-Leduc *et al*. Frequency estimation based on the cumulated Lomb–Scargle periodogram. *Journal of Time Series Analysis*, 29(6):1104–1131, 2008.
[38] Constantino A. García *et al*. A new algorithm for wavelet-based heart rate variability analysis. *Biomedical Signal Processing and Control*, 8(6):542–550, 2013.
[39] Munenori Fukunishi, Daniel Mcduff, and Norimichi Tsumura. Improvements in remote video based estimation of heart rate variability using the Welch FFT method. *Artificial Life and Robotics*, 23(1):15–22, 2018.
[40] Arnaud Delorme and Scott Makeig. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.
[41] Irene Winkler *et al*. Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, 11(3):035013, 2014.
[42] Richard J. Davidson. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition and Emotion*, 12(3):307–330, 1998.
[43] Rafael Ramirez *et al*. Musical neurofeedback for treating depression in elderly people. *Frontiers in Neuroscience*, 9:354, 2015.
[44] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3):489–501, 2006.
[45] Armando M. Oliveira *et al*. Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. Proceedings of Fechner Day, 22:245–250, 2006.
[46] Michel Valstar *et al*. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. Proceedings of the 6th International Workshop on Audio/visual Emotion Challenge, 3–10, 2016.
[47] Yuhei Ikeda, Ryota Horie, and Midori Sugaya. Estimating emotion with biological information for robot interaction. *Procedia Computer Science*, 112:1589–1600, 2017.
[48] Maro G. Machizawa *et al*. Quantification of anticipation of excitement with three-axial model of emotion with EEG. *bioRxiv*, 659979, 2019.