Stress Estimation Using Multimodal Biosignal Information from RGB Facial Video

Takumi Nagasawa
Chiba University
Chiba, Japan
takumin72orange@gmail.com

Ryo Takahashi
Chiba University,
Chiba, Japan
takahashi0705@chiba-u.jp

Chawan Koopipat
Chulalongkorn University,
Thailand
tsumura@faculty.chiba-u.jp

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

Abstract

In the present paper, we propose a method for acquiring multiple biological information inputs from a red-green-blue (RGB) facial video footage and using their feature values to estimate stress levels. Such estimations are important because if left unchecked, stress can cause severe mental illness and/or physical damage to the human body. Accordingly, it is important to understand the onset of stress at an early stage and take measures to counteract it. However, since it is difficult for us to accurately gauge our stress levels, it would be desirable to establish an objective and accurate estimation method. Additionally, while the commonly used questionnaire method is easy to implement, it lacks both objectivity and accuracy. In a recent study, many methods that use biological information were proposed. In the present study, we estimate stress using three biological signals captured using an RGB camera: pulse, blinking rate, and pupil diameter. Our results show that stress estimation accuracy is improved by using these biological signals, thereby indicating that it is possible to estimate stress more accurately by using biological information in a multimodal manner.

1. Introduction

It is necessary to quickly and accurately grasp stress levels and take appropriate remedial actions because unchecked stress can cause enormous mental and physical damage. Current stress estimation methods are generally based on questionnaires, but it is not always possible to perform objective and accurate stress estimations with such methods because they depend on the subjective evaluations of the respondents.

In previous studies, methods that use changes in biological information such as pulse and pupil dilation have been used to estimate stress levels more accurately. A particular advantage of methods using biometric information is that measurements can be continuously collected without the subjective biases of the subjects. Such measurements are normally collected using contact-based dedicated equipment, but noncontact measurement methods are more desirable because contact-based methods have limited measurement environments.

In a previous study by Mitsuhashi *et al.* [1], pulse waveforms were extracted by skin pigment component separation from a facial video captured using red-green-blue (RGB) camera footage, after which four-stage stress classification was performed using the features obtained from the pulse waveforms. This study aims to build on that method, thus increasing the accuracy of stress classification, by extracting noncontact information on measurable blinking and pupil dilation from facial videos, in addition to extracting pulse waveform data.

2. Proposed method

It is widely known that stress affects the dominance and autonomic balance of the nervous system and that those effects are partially expressed as physiological changes such as increased heart rate and pupil dilation. Because these changes occur unconsciously, objective observations of autonomic nervous system activities can be made by measuring changes in such biological information. In this study, we acquire and utilize pulse, blinking rate, and pupil dilation as biosignal data extracted from facial video footage captured by an RGB camera, and then use those extract features for stress estimation.

2.1 Pulse waveform estimation

A pulse is defined as a change in blood pressure or volume in the peripheral vasculature that accompanies the heartbeat. A contact-type measurement method using a machine called a photoplethysmograph is generally used when measuring the pulse for medical purposes. Using the property that hemoglobin contained in the blood absorbs green light, this machine obtains pulse waveform data by observing the intensity of light emitted from and reflected to a fingertip attachment. However, this method requires special equipment, which reduces the environments in which it can be used, and can also increase stress due to the need for the person being tested to make physical contact with the device.

Therefore, a technique called image-based photoplethysmography (iPPG) has been proposed as a noncontact method for acquiring pulse waveforms. For example, Verkruysse *et al.* [2] proposed a method for acquiring pulse waveforms that works by calculating temporal changes in average pixel values of G signals of region of interest (ROI). This method utilizes the fact that the G component of the skin in the moving image captured by the RGB camera correlates with the blood volume, which means it can be used to acquire the pulse waveform from time-series changes to the G component. However, since this method is affected by lighting fluctuations and body movements (which increase the observed noise), several studies have been conducted to find ways of acquiring clear pulse waveform data from facial video footage [3-6].

Fukunishi *et al.* [7] proposed a method of measuring pulse waveforms that is robust against lighting fluctuations based on a technique called skin pigment component separation. In this method, hemoglobin components are extracted by applying skin pigment component separation to captured skin video footage [8]. This method estimates the pulse waveform by calculating the average pixel value in the skin area of the face using the hemoglobin component of the video that fluctuates with heartbeat blood volume changes. Furthermore, a more accurate pulse waveform can be obtained by applying a band-pass filter to extract and apply a band of 45 to 180 beats per minute (the normal range of a human heart) to the estimated pulse waveform. Figure 1 shows time-series changes of a pulse waveform obtained from a facial video.

2.2 Blinking estimation

It has been noted previously that blinking rates are affected by emotional stimulation levels, such as tension and anxiety. It is also known that the sympathetic nerve controls a portion of the eyelid muscles that perform blinking activities. In general, when measuring blinking activity, a method of acquiring an electromyogram (EOG) of the eye using an electromyograph is widely used. However, since this is also a contact-style measurement, a noncontact measurement method is needed to facilitate stress-free measurements.

Soukupová and Čech [9] proposed a method that detects eye blinks from facial video footage by using landmarks. In the method described above, 68 landmarks are acquired from a facial video by machine learning, and the degree of eye-opening/closing is determined based on the vertical and horizontal distance fluctuations of six points related to the eyes. Figure 2 shows an example of the six landmarks acquired and the related vertical and horizontal distance fluctuations, while Equation (1) shows the calculation formula of the biological signal for a blink.

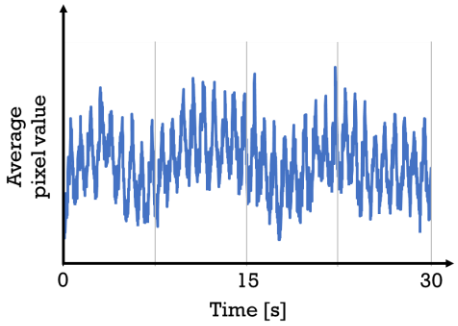

Figure 1. Pulse waveform from RGB facial video.

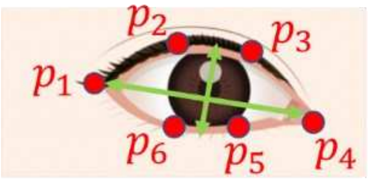

Figure 2. Changes to eye landmark distances.

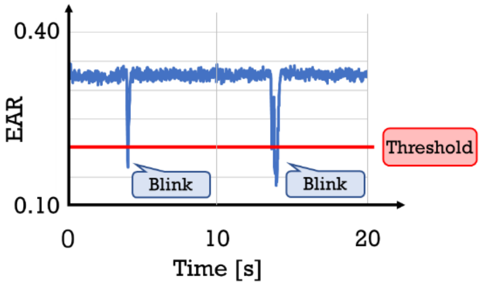

Figure 3. The blink waveform calculated by Equation (1).

$$\mathrm{EAR} = \frac{p_2p_6 + p_3p_5}{2(p_1p_4)} = \frac{\textbf{The Average of Lateral Length of Eye}}{\textbf{The Average of Vertical Length of Eye}} \qquad (1)$$

The eye aspect ratio (EAR) calculated by Eq. (1) represents the ratio of the vertical and horizontal distance of the eye based on the six landmarks mentioned above. While the horizontal distance of the eye does not change when the eye is opened or closed and, the vertical distance is kept constant when the eye is open, the length decreases rapidly when the eye is closed, such as when a blink occurs.

Figure 3 shows the blink waveform, which is the time-series change of the EAR value calculated by Eq. (1). The part where the value decreases rapidly is considered to be the blinking point. From the blink waveform seen in Fig. 3, a total of 14 features, such as the number of blinks, the amplitude at eye opening and closing times, and the opening and closing speeds, can be acquired.

2.3 Pupil estimation

In this subsection, we explain how pupil diameter changes are acquired using circles fitted to the pupil and iris that have been segmented from a facial video. Basically, pupillary movement and size are related to the functions of the pupillary sphincter and pupillary dilator muscles. They are controlled by the autonomic nervous system, and it is known that fluctuations in pupil diameter can be affected by stress. Specifically, the pupillary sphincter is under the control of the parasympathetic nerve and the pupillary dilator muscles are under the dual control of the sympathetic and parasympathetic nerves. Thus, the pupil shrinks in the resting state and enlarges in the stressed state, which means it can be a useful stress indicator.

In this study, the eye region is extracted from the facial video, and circles are fitted to the pupil and the iris segmented by deep learning. The ratio of each size is then used to obtain pupil diameter changes. Figure 4 shows the procedure for acquiring the pupil and iris sizes from the facial video.

Segmentation of the pupil/iris region from the facial video footage is performed using the U-Net deep learning network, which is a full-layer convolutional network specialized for image segmentation. U-net was selected because it is capable of using fewer images for learning, can be trained quickly, and has high image segmentation accuracy. The pupil/iris ratio (PIR) is calculated by comparing the diameters of the circles fitted in Fig. 4. This value is important because while the iris size does not change due to stress, the pupil size varies. Equation (2) is an equation for calculating the PIR.

$$\mathrm{PIR} = \frac{a_p}{a_i} = \frac{\textbf{Diameter of Circle Fitted to Pupil}}{\textbf{Diameter of Circle Fitted to Iris}} \tag{2}$$

By using the ratio of the pupil to the iris as an index, pupil fluctuation can be observed regardless of the measured distance. Figure 5 shows an example of a PIR time-series change calculated by Equation (2).

Referring to Fig. 5, a sharp decrease similar to the blink waveform shown in Fig. 3 is observed. This is thought to be because the pupil could not be detected due to blinking, which means the numerical value was lost. To continuously analyze pupil fluctuations, Hermitian interpolation is performed to compensate for the loss of the numerical value. Figure 6 shows the result of performing interpolation on the PIR values provided in Fig. 5. As shown in Fig. 6, it is possible to eliminate extreme changes in the PIR numerical value by performing numerical interpolation.

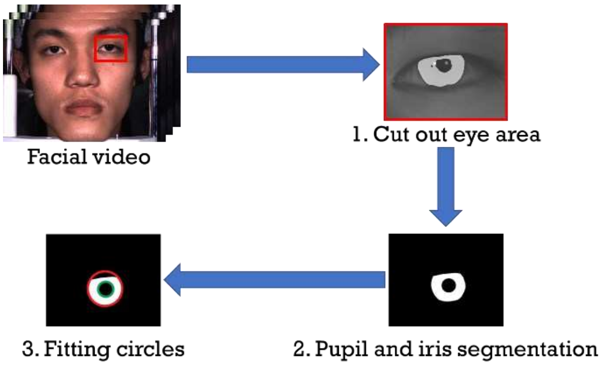

Figure 4. Procedure for acquiring the pupil and iris size from the facial video.

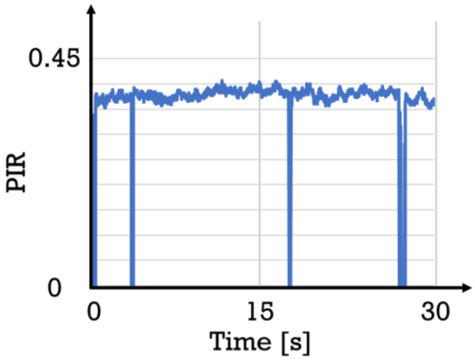

Figure 5. Example of PIR time-series changes.

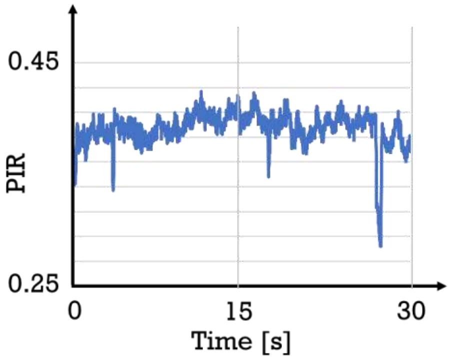

Figure 6. The result of performing interpolation on the PIR values in Figure 5.

3. Feature extraction

In this research, as part of efforts to improve stress estimation accuracy levels, we obtain and then combine the multimodal features of the three types of biological information described in the previous section. In this section, we explain the features used in this research.

3.1 Pulse features

It is known that changes in biological information due to stress appear as fluctuations in the interval between peaks of the pulse waveform. Since pulse rate variability is calculated by noting differences between adjacent peak times of the pulse waveform, we first calculate the pulse rate variability as shown in Fig. 7. We then obtain other pulse features by analyzing the pulse waveform in the time and frequency domains.

3.1.1 Time-domain analysis

The time-domain features, which can be easily obtained by the pulse waveform interval shown in Fig. 7 are analyzed directly. In particular, the value of the pulse waveform interval $meanRRI$ and the standard deviation $stdRRI$, as well as the average value of the heart rate waveform $meanHR$ and the standard deviation $stdHR$, all of which are calculated from one pulse waveform interval, are the most easily obtainable indices.

Among these, the standard deviation $stdRRI$ of the pulse waveform interval reflects the overall fluctuation of the pulse waveform, while the root mean square $RMSSD$ of the pulse waveform sequential difference reflects short-term pulse waveform fluctuations. The method used for calculating $RMSSD$ is shown in Eq. (3).

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}\left(RR_{j+1} - RR_j\right)^2} \qquad (3)$$

In this equation, N presents the total number of consecutive pulse waveform intervals while RR_j refers to the jth pulse waveform interval. In the time domain, we extract the five abovementioned features.

3.1.2 Frequency Domain Analysis

Features in the frequency domain are obtained by analyzing the power spectral density (PSD) of the pulse interval. In this study, the PSD is calculated using the periodogram proposed by Lomb and Scargle. Because the changes to the high-frequency (HF: 0.15 to 0.40 Hz) and low-frequency (LF: 0.04 to 0.15 Hz) components of the pulse waveform fluctuations reflect the actions of the autonomic nervous system, our method also uses the features obtained from the frequency domain analysis of the pulse waveform.

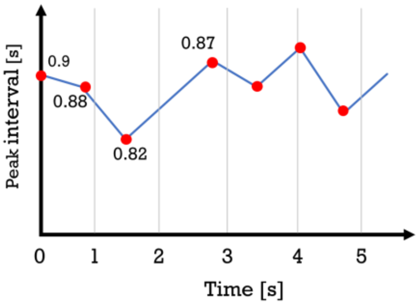

Figure 7. Pulse rate variability.

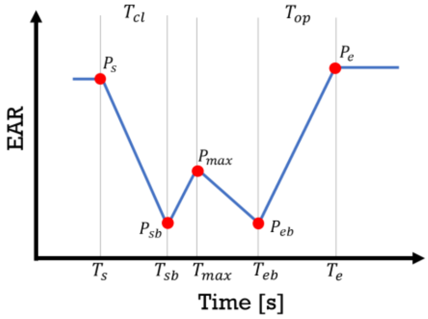

Figure 8. Blink feature points.

In this research, we calculate LF as the integrated values of the LF band in the PSD calculated by the Lomb-Scargle periodogram. In addition, HF is calculated as the integrated values of the HF band, nLF is calculated as the normalized LF, nHF is calculated as the normalized HF, and LF/HF is calculated as the ratio of LF per HF.

As described above, we acquire a total of ten feature types from the pulse waveform.

3.2 Blink features

Next, we will explain the extraction of features which are related to blinks. Figure 8 schematically shows the blink waveform described in Section 2 and the feature points obtained from it. In Fig. 8, P_s and P_e represent the start and endpoint of the blink, respectively, while P_{sb} and P_{eb} represent the endpoint of the eyelid closing process and the start point of the eyelid opening process. P_{max} represents the time when the EAR value reaches a maximum. Based on each of these feature points, the 10 feature values can be extracted.

First, the amplitude at the time of eyelid closure A_{cl} is defined from the difference between the EAR values of P_s and P_{sb}. Next, the amplitude A_{op} at the time of eyelid opening is determined from the difference between the EAR values of P_{eb} and P_e. The maximum amplitude A_{mv} is then defined from the difference between the average value of P_{sb} and P_{eb} and the EAR value of P_{max}. The eyelid speed $V_{cl} = A_{cl}/T_{cl}$ is defined from the time difference $T_{cl}(=T_{sb}-T_s)$ between P_s and P_{sb}, while the eyelid opening is defined from the time difference $T_{op}(=T_e-T_{eb})$ between P_{eb} and P_e the speed $V_{op}=A_{op}/T_{op}$.

The mean and standard devaition values of these parameters are obtained as feature values. Additionally, we use the number of blinks EB_num_{all} when P_{sb} and P_{eb} are detected as individual peak detection points, the number of blinks counted from P_s to P_e as one blink EB_num_{ecp}, and Eye_Closed_Time, which is the sum of the time of closed eyes from P_s to P_e. Thus, a total of 14 feature types can be obtained from the blink waveform.

3.3 Pupil features

In this subsection, we will explain the extraction of features related to pupils. As shown in Figs. 5 and 6, we acquire a total of pupil six feature value types. These are the $meanPIR_{ori}$, which is the average of the original PIR values seen in Fig. 5; $stdPIR_{ori}$, which is the standard deviation of the original PIR values; $meanPIR_{ip}$, which is the average of the interpolation values shown in Fig. 6; $stdPIR_{ip}$, which is the standard deviation of the interpolated PIR values; $minPIR$, which is the minimum PIR value; and $maxPIR$, which is the maximum PIR value.

4. Experiment

Using the biological signals described in the previous sections and their obtained features, we measured four stress state levels and then verified the stress classifications and their accuracy using a multiple regression analysis-based classifier.

Figures 9 and 10 show the experimental setting and the procedures used. Our experiment was conducted in a dark room with nine students in their 20s (six male and three female) participating as test subjects. Video footage was captured using an RGB camera with a frame rate of 60 fps and a resolution of 1024×768 pixels. Artificial sunlight was provided as lighting.

In these experiments, each test subject was asked to use a chin rest to stabilize his or her face in order to prevent body movements from being introduced into the pulse waveform information as noise. Two minutes of video footage were taken for each of the subjects in four stress level states (a relaxed state and three states in which they were required to solve mental arithmetic tasks of varying difficulty).

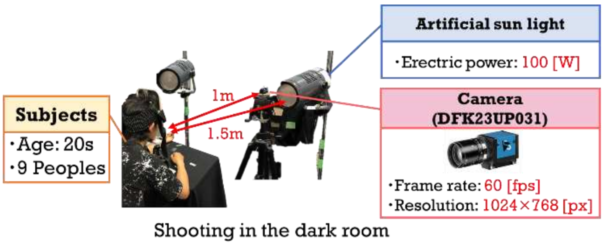

Figure 9. Experimental setting.

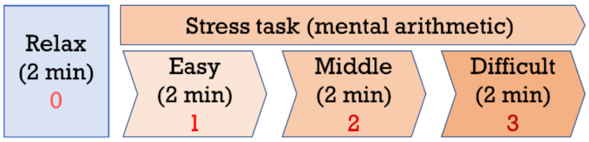

Figure 10. Experimental procedure.

A six-minute rest interval was set between video captures to eliminate the effects of the previous task. The mental arithmetic tasks given to the participants were divided into three levels. For the easiest task, they were asked to perform multiplication with two single-digit numbers. For the moderate level task, they were asked to perform multiplication with one double-digit number and one single-digit number. For the most difficult task, they were asked to perform multiplication with two double-digit numbers.

During the video capturing process, the subjects were instructed to perform each task while looking straight at the camera and keep their eyes open as much as possible so that we could record blink and pupil information accurately. To subjectively evaluate the degree of stress experienced by the test subject in each task, a post-evaluation was performed using the State-Trait Anxiety Inventory (STAI) questionnaire, which makes it possible to evaluate the anxiety state a subject feels transiently during a specific scene (such as during the execution of a stress task).

5. Results

In the present study, we compared classification accuracies by using features obtained from facial videos, which were combined for each biological signal (for example, pulse only, blink only, pulse and blink, all biosignals) and then performed multiple regression analyses for each combination. Five features that were selected based on Pearson's product-moment correlation coefficient for each combination of biological signals were used for the stress estimations. We also performed cross-validation by the leave-one-out method to verify the stress estimation accuracy.

To estimate stress levels, we created a model formula by performing multiple regression analysis on teacher data and then calculated a predictive label value by inputting the test data. A feature of the multiple regression analysis estimation method is that we can calculate the estimation value as a constant value, which means that it may be possible to estimate a change in a stress state time-series in the future.

As an indicator of classification accuracy, we compare the average value of the difference between the correct label and the prediction value by the root mean square error (RMSE). Specifically, we express the correct label by discretely assigning numerical values to each state where the video was taken, as shown with red letters in Fig. 8. Equation (4) expresses the calculation formula of the RMSE for the correct answer label and the predicted value.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (4)$$

In this equation, N presents the number of data, y_i presents the prediction value, and $\hat{y}_i$ refers to the correct label value. A lower RMSE value indicates a more accurate estimation.

Figure 11 shows the RMSE for each combination of biometric information. We also compared the coefficient of determination of the multiple regression analysis performed using all data. The coefficient of determination is an index indicating the goodness of fit of the model formula to the data. It is considered that the higher the coefficient is, the better the model fits the data.

Figure 12 shows the coefficient for each biometric information combination. From Figs. 11 and 12, we found that estimation accuracy improved as the number of biological information indicators used increased and that the estimation accuracy was maximized by combining information on pulse, blinking rate, and pupil dilation values.

Table 1 shows the feature values selected by the combination using all of the biological information collected in this study, and the effectiveness of multimodal use of biometric information in stress estimations.

It is worth noting that LF/HF, which is known to be an effective index for stress estimation, was not selected as a feature. We believe that factors other than the stress task, including the environmental load imposed by taking video footage in a dark room while illuminating the test subject's face with a bright light, and the long-term stress associated with the daily mental state of the test subject, should also be considered. In order to eliminate the effects of these factors, the following countermeasures can be considered.

The environmental stress imposed during the video captures in this study can be prevented by providing sufficient time for the test subject to adapt to the environment before conducting the measurements. We also believe that long-term stress can be mitigated by providing time before taking measurements in the same way as with environmental factors. However, it would be difficult to completely eliminate the stress effects caused by the experiment setting.

Feature values	Biological information
$\boldsymbol{stdHR}$	Pulse
$\boldsymbol{stdRRI}$	Pulse
$\boldsymbol{EB\_num_{ecp}}$	Blink
$\boldsymbol{Eye\_Closed\_Time}$	Blink
$\boldsymbol{minPupil}$	Pupil

Table 1. Feature values selected by the combination using all biological information.

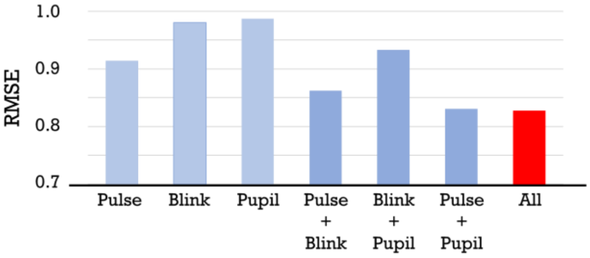

Figure 11. RMSE for each combination of biometric information.

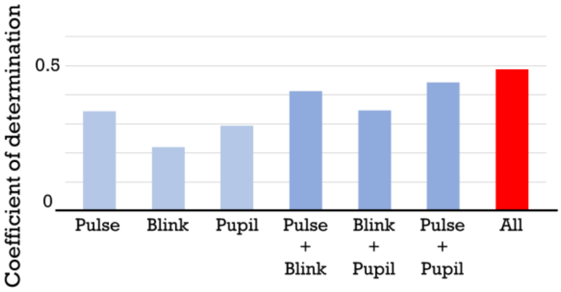

Figure 12. Coefficient for each combination of biometric information.

Looking at the feature value changes between the states, we can see a significant difference between the resting state and the stress task, especially in the blinking information such as EB_num_{all}. Since this may provide a useful index for determining such short-term stress, it will be necessary to conduct further experiments in the future for clarification purposes.

6. Conclusion

In this paper, we proposed a method for acquiring three types of biological information from captured RGB facial video footage and used that data to estimate stress levels. Through experiments in which we combined these multiple biometric information types, we confirmed that the stress estimation accuracy was improved.

In future studies, by applying appropriate stress to the test subjects, we will perform measurements to verify the effectiveness of this method. We will also consider applying this method to more complex stress situations. Additionally, in order to analyze pupil information in more detail, such as we currently do with pulse waveforms, it will be necessary to consider methods for analyzing pupil changes.

References

[1] Ryota Mitsuhashi, Kaito Iuchi, Takashi Goto, Akira Matsubara, Takahiro Hirayama, Hideki Hashizume, and Norimichi Tsumura. Video-based stress level measurement using imaging photoplethysmography. *2019* IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 90–95, 2019.

[2] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. Optics Express, 16(26):21434-21445, 2008.

[3] Y. Yang, C. Liu, H. Yu, D. Shao, F. Tsow, and N. Tao. Motion robust remote photoplethysmography in CIELab color space. Journal of biomedical optics, 21(11):117001, 2016.

[4] D. Mcduff, S. Gontarek, and R. Picard. Improvements in remote cardiopulmonary measurement using five band digital camera. IEEE Transaction on biomedical engineering, 61(10), 2014.

[5] M. Poh, D. Mcduff, and R. Picard. Advancements in noncontact, multiparameter physiological measurement using a webcam. IEEE Transaction on biomedical engineering, 58(1), 2011.

[6] D. Shao, Y. Yang, C. Liu, F. Tsow, H. Yu, and N. Tao. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. IEEE Transaction on Biomedical Engineering, 61(11):2760-2767, 2014.

[7] M. Fukunishi, K. Kurita, S. Yamamoto, and N. Tsumura. Non-contact video-based estimation of heart rate variability spectrogram from hemoglobin composition. Artificial Life and Robotics, Vol. 22, No.4:457-463, Springer, 2017.

[8] N. Tsumura, N. Ojima, K. Sato, M.Shiraishi, H. Shimizu, H.Nabeshima, S. Akazaki, K. Hori, and Y. Miyake. Image-based skin color and texture analysis / synthesis by extracting hemoglobin and melanin information in the skin. ACM Transactions on Graphics (TOG), 22:770-779, 2003.

[9] T. Soukupová, and J. Čech, Real-Time Eye Blink Detection using Facial Landmarks. 21st Computer Vision Winter Workshop, 2016.