High-dimensional biomedical image recognition with artificial intelligence

by

Kazutoshi Ukai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Engineering

Supervisor: Professor Syoji Kobashi

University of Hyogo, Japan

March 2022

Preface

Advances in biomedical imaging technology have made it possible to acquire large amounts of high-dimensional image data such as video images, 3D volume data, and fusion images of multiple modalities, consequently increasing the need for automation. In recent years, image recognition technologies that surpass humans have been developed, such as face recognition technology that distinguishes slight individual differences in the face, and skin lesion detection technology that has higher diagnostic accuracy than specialists. It is expected to be applied to the video images and 3D volume data, and expand the range of application. Although it has become possible to distinguish individual differences in the face, research on grasping the internal condition of an individual from the complexion and facial expression is still under development. Also, in lesion detection from medical images, the accuracy is increasing by utilizing deep learning. However, the calculation cost in applying deep learning to a 3D volume data is still one of the major concerns. This thesis introduces methods to solve these problems.

Firstly, a study is conducted focusing on blood flow as a biomarker for recognizing the internal state from facial video images. By calculating the cycle of each heartbeat, which is called R-R interval (RRI), from the blood flow, the state of the autonomic nervous system can be quantified based on the fluctuation of RRI. In recent years, some studies have been conducted to extract the information of blood flow from facial video images. However, it takes an observation time of several tens of seconds to extract the signal component, and it is a serious problem when used in real-time applications. This study enables to measure RRI accurately with a short observation time by simultaneously observing multiple regions on the facial video images.

Secondly, a method is introduced to detect spoofing by a 3D artificial object, which is a problem in face recognition. It is based on blood flow which is detected from facial video images. By analyzing the blood flow change in the facial video image in multiple wavelengths and multiple regions, impersonation was detected in the proposed method for three kinds of spoofing objects - 3D face models, printed photographs, and still images of monitors.

Thirdly, a method is proposed for detecting the 3D pelvic fractures from CT images. Deep convolutional neural networks were used to detect 2D fracture candidates from

multi-directional 2D cross-sectional images reconstructed from 3D-CT volume, and the detected 2D fracture candidates were integrated to obtain 3D fracture candidate regions. By analyzing 3D-CT volume in 2D space, the calculation cost could be significantly reduced. Also, because deep neural networks were trained by using multi-directional 2D cross-sectional images, a large number of 2D images could be synthesized from each 3D-CT volume.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor Professor Syoji Kobashi of University of Hyogo for the support of my Ph. D study with his immense knowledge. I also sincerely thank to Professor Naotake Kamiura, Professor Shinsaku Hiura, Assistant Professor Daisuke Fujita, Researcher Rashedur Rahman (University of Hyogo), and Associate Professor Naomi Yagi (Himeji Dokkyo University), for their kind support. I also thank to all members of Data Science Laboratory of University of Hyogo for cooperating in every phase of this work.

My research would not be completed without medical support and advice. I would like to express my deep gratitude to Dr. Hirotsugu Muratsu, Vice President, Dr. Akihiro Maruo, Manager of Ortho-Plastic Trauma Center, and Mr. Keigo Hayashi, Radiological Technologist, Steel Memorial Hirohata Hospital. And also, I am extremely grateful to GLORY LTD. for giving the opportunity to going on to Ph. D study. Especially, I would like to sincerely thank to Mr. Motozumi Miwa, President, Dr. Hirofumi Kameyama, Senior General Manager of R&D Center, and Mr. Kazuo Fukumoto, General Manager of New Technologies Creation Dept., GLORY LTD., for their consideration and support in various aspects. Moreover, I sincerely thank to Dr. Hiroyuki Onishi, Dr. Toru Yonezawa, and Mr. Hisakazu Yanagiuchi of Glory Industry-Academia Joint Laboratory for Medical Engineering for their kind cooperation.

Finally, I am grateful to my friends who gave many encouragements and praises to my challenge. Also, I would like to thank for the encouragement of my sons and my wife, and the watching from behind of my parents.

Contents

Chapter 1 : Introduction

Recent rapid advances in biomedical imaging technology has made it possible to acquire a large amount of high-dimensional image data, such as moving images, three-dimensional (3D) volume data, and fusing images of multiple modalities. As a result, it has become difficult for humans to adequately respond to the needs of medical, security, and other applications. In addition, technologies that outperform humans have been developed for many tasks including the tasks that require advanced human cognitive abilities that could not be replaced by artificial intelligence in the past. Also the need for automation of biomedical image recognition is increasing. One of the examples is face recognition technology that can distinguish slight individual differences in faces [1]. Another example is in the field of medical image recognition, where technology with diagnostic accuracy that exceeds the performance of medical specialists in detecting skin diseases [2]. Although it has become possible to distinguish individual differences in faces, research on understanding the internal state of individuals from the facial color and expressions is still at its infancy [3]. In addition, the accuracy of lesion detection from medical images has been increasing with the use of deep learning, but issues such as the increase in computational cost in applying it to 3D volume data remains [4].

Chapter 2 proposes an estimation method of R-R interval (RRI) from facial video images [5]. In this study, blood flow is considered as a cue for detecting internal states from face images. By calculating RRI from blood flow, we can quantify the state of the autonomous nervous system based on the variation of RRI [6][7][8]. To measure blood flow, a photoplethysmographic device that utilizes the change in light absorption due to the increase or decrease in blood volume associated with the heartbeat [9] was developed several decades ago. In recent years, a new approach based on the same principal has been conducted to detect blood flow from facial video images taken with a commercial color camera [10]. However, since many noise components are included on facial video images, advanced processing such as independent component analysis (ICA) is required. Therefore, it takes longer observation time to extract the signal components, which is inconvenient when it is used for real-time applications [11][12]. In this study, a novel method to measure RRI accurately is proposed with a short observation time by simultaneously observing

multiple regions on the facial video image instead of the length of observation time.

Chapter 3 investigates an application of obtained blood flow information from facial videos to detect impersonation by 3D objects [13], which is a problem in face recognition. It proposes a method to detect impersonation by four kinds of spoofing objects such as a 3D face model, a photograph, a still image displayed on a screen, and a moving image by analyzing the blood flow changes in the face video image in multiple wavelengths and regions.

Chapter 4 proposes an automated method for detecting 3D lesions in CT images, which can be applied to 3D medical images such as CT and MRI images [14]. The main issues regarding applying deep learning to 3D medical images are small number of subjects, very large size of data, high computational cost, etc. [4]. In this study, these issues are addressed, and we propose a novel method for detecting pelvic fractures on CT images.

This dissertation consists of five chapters, including this chapter. In Chapter 2, a multiple-measurement-points-voting-method is proposed to estimate RRI from time-series face images in a short time. The experimental results are compared with the ICA method to describe the effectiveness of the proposed method. In Chapter 3, a liveness detection method is proposed based on changes in blood flow on facial skin using video images, and it is evaluated using 16 subjects in five types of lighting environments and four types of spoofing objects including 3D objects. In Chapter 4, a method is proposed for detecting pelvic fractures from 3D-CT using a deep convolutional neural network with multi-directional slab images, and it is evaluated using 93 subjects with fractures and 112 subjects without fracture. Chapter 5 summarizes this dissertation.

Chapter 2 : Short-time estimation of R-R interval from facial video image with a multiple-measurement-points-voting-method

2.1 Introduction

In the modern society, mental stress accumulates without people being aware of it due to long hours of work, especially using electronic displays in the workplace. Since mental stress affects the endocrine system and autonomic nervous system, many studies have been conducted to quantify and analyze stress. Among them, a relationship between stress and heart rate variability has attracted much attention [6][7][8]. Here, heart rate variability refers to the variation of the period (RRI) of each heartbeat in a pulse wave.

The pulse wave can be obtained using an electrocardiogram or, for simplicity, a photoplethysmographic device that detects the pulse wave from the finger or earlobe. In recent years, wearable devices such as wristwatches and shirts have been developed, making constant observation possible to the pulse wave [15]. However, these devices are worn consciously by the subject and are not suitable for observing stress that accumulates unconsciously. On the other hand, measurement with millimeter-wave sensors [16] has been proposed as a non-contact method, but it requires special and expensive equipment.

Therefore, in this study, we focus on face video images captured by a video camera. Since the face is a part of the body that almost everyone is exposed to in daily life, its applicability is broad, and the camera equipment is inexpensive and easy to install. However, if we assume that the video images will be used in actual living and office environments, we need an analysis method that can measure the heart rate in a shorter observation time.

Some methods using frequency spectrum analysis and ICA to calculate the heart rate from facial video images have been proposed [10][11][12][17][18]. Takano *et al.* proposed a method to estimate the heart rate by using a band-pass filter and frequency spectrum analysis on the average time-series signal of the entire face region [17]. Jiang *et al.* applied Kalman filter to remove white noise and estimated the heart rate by frequency spectrum analysis on the obtained time-series signal [18].

In order to apply frequency spectrum analysis, they require an observation time of about 30 sec. On the other hand, ICA-based methods have been proposed in the literature [10][11][12]. These methods do not use frequency spectrum, but they require a certain observation time for signal decomposition with ICA.

In this study, we propose a method that detects the pulse wave peak from the time-series signal of each pixel. The pulse wave peaks are extracted using a differential filter, and then it finds the most frequent value of the RRI in every pixels of region-of-interest (ROI). Hereinafter, the proposed method is called as multiple-measurement-points-voting-method. This method is based on a signal filter without using the frequency spectrum or ICA, and thus it can shorten the observation time. In addition, by using a large number of observation points simultaneously, the RRI can be estimated robustly and quickly by suppressing the influence of noise from non-skin parts such as spectacles and hair.

The details of the method are presented in the following sections; the subjects and measurement data used in the evaluation experiments of this study are presented in section 2.2; the proposed multiple-measurement-points-voting-method is described in section 2.3; the experimental results are discussed in section 2.4; a summary is given in section 2.5.

2.2 Experimental data

The experimental data were collected in two different states: stationary state and working state while working on a computer. This study was conducted by following procedures approved by the Ethics Committee of the University of Hyogo on "Estimation of Physical and Mental States Using Time-Series Facial Images", with the obtained written consent from all subjects.

2.2.1 Stationary state

The experimental data at the stationary state were collected from four males and one female, twice a day (around 12:00 and 17:00) for 3 minutes each time. A total of 198 data were collected (this dataset is referred to as DS1). To evaluate the robustness within subjects, a large number of data were conducted with a small number of subjects. Table 2.1 shows the demographic of subjects (gender, age, and whether or not they wore spectacles) and the number of trials for each subject. None of the subjects wore makeup.

Table 2.1 Demographic of subjects

Subject	Sex	Age (y.o.)	Spectacles (Type)	Number of trials	
				Noon	Evening
A	M	23	No	11	20
B	M	22	Yes (Black edge)	16	26
C	M	22	Yes (Borderless)	31	30
D	M	25	No	27	27
E	F	22	No	6	4

Next, in order to evaluate the robustness between subjects, data were collected from 21 males and 6 females. Two data were collected from each subject, one with spectacles and one without spectacles. The ages of subjects ranged from 21 years to 62 years with a mean age of 30.2 years and a standard deviation (SD) of 13.6 years. All of the subjects were different from the subjects in DS1. The total number of collected data was 54 (this dataset is referred to as DS2). Spectacles without power were used, and the women wore make-up for daily outings.

Face video images were collected by using a color camera (Allied Vision Mako 192C, 1600×1200 resolution, 40fps, 12mm lens), which was synchronized with a fingertip photoplethysmographic device (Tokyo Devices IWS920, 409.6 [sample/s] sampling rate). The video images captured by the color camera were stored as a lossless compressed video where each pixel had three channels: red (R), green (G), and blue (B), with a color depth of 8 bits per channel. Each video was taken in 3 min.

The data acquisition setup is illustrated in Fig. 2.1. The distance between the camera and the subjects was fixed at 1 m. The height of the subjects were adjusted by changing the height of chair so that the entire face was captured. The videos were acquired in a room under general lighting conditions. During the video acquisition, the subjects wore a photoplethysmographic device on their finger. Also, the subjects were instructed to look at the camera with open eyes in a relaxed state.

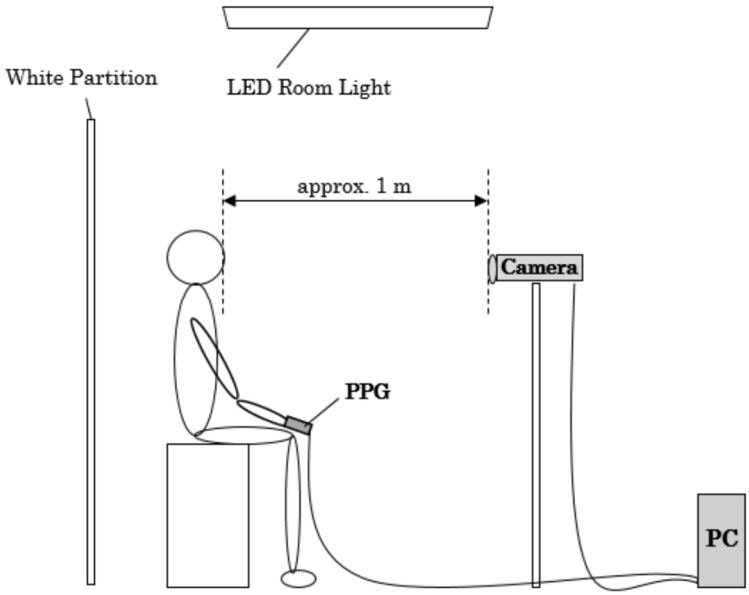


Fig. 2.1 Data acquisition booth.

2.2.2 Working state

Six videos were acquired from eight males and two females, whose age was between 26 and 57 years (mean±SD of 38.4 ±11.1 years). In total, 60 data for working state were acquired (this dataset is referred to as DS3.). The females wore makeup for daily outings.

The face video images were acquired by using a color camera (Basler acA1920-155uc, 1920×1200 resolution, 30fps, 8mm lens), which was synchronized with a photoplethysmographic device (Nihon Kohden OLV-4202, sampling rate 1000 [sample/s]). When the videos were collected while the subjects were working in computers, the photoplethysmogram (PPG) signals were acquired from the earlobe. The stored videos had the same configuration as the videos stored in the stationary state.

The subject sat at a desk with a camera fixed on two 20-inch displays (the distance from the camera to the subject's face was about 0.5 m) as shown in Fig. 2.2, and facial videos were taken while the subject was working on a computer. A photoplethysmographic device was attached to the earlobe, and the subject was instructed not to touch his face with his hands during the recording. The recording was conducted under a general indoor lighting environment.

DS3 were collected by using an in-house experimental program. The program moved a pointer on a screen at an irregular speed (maximum speed of about 90 cm/s), and the subject was instructed to follow the movement of the pointer. Then the

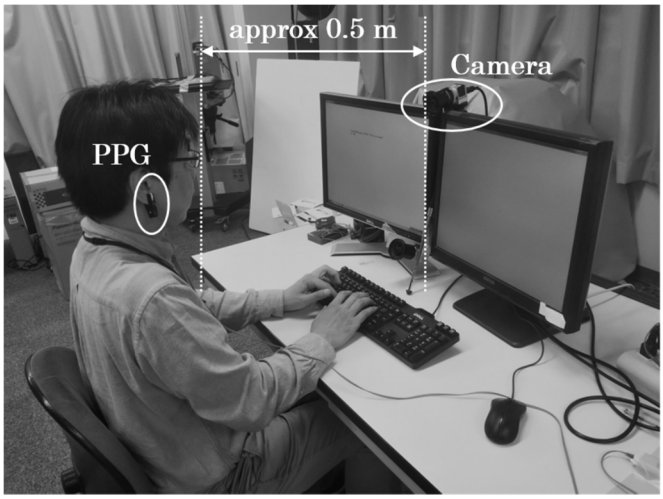


Fig. 2.2 Data acquisition environment.

subject typed an alphabetic string of about 20 characters displayed by the program near the pointer. The duration of each video was two minutes. During the acquisition, the subjects were instructed to repeat two tasks consecutively - following the mouse pointer on screen and typing, each task for 15 sec. The goal of these experiments was to simulate a typical work environment on a computer.

2.3 Multiple-measurement-points-voting-method

2.3.1 Overview

The multiple-measurement-points-voting-method estimates an RRI for each pixel in the ROI shown in Fig. 2.3 (A: face or B: cheek). The RRI is estimated from the time difference between two consecutive peaks detected by the first order gaussian derivative of the normalized time-series signal. Then, the most frequent value of the RRI calculated at every pixel in the ROI is found as the RRI at that time. The details of the proposed method are described below.

2.3.2 Extraction of regions of interest

This method utilizes threes ROIs, the face ROI, the cheek ROI, and the nose ROI. Each ROI is determined by detecting facial feature points (dots in Fig. 2.3) on the first frame of the video using the C++ library named Dlib [19][20]. The detected nose ROI is used as the template image, and the following frames are aligned by tracking the template in order to extract the face ROI and the cheek ROI. The relative positions of the face ROI from the nose ROI are obtained. The likelihood of template matching used is correlation coefficient. Next, as a preprocessing, each channel of

RGB image of the video is spatially smoothed using a moving average filter of $S \times S$ pixels.

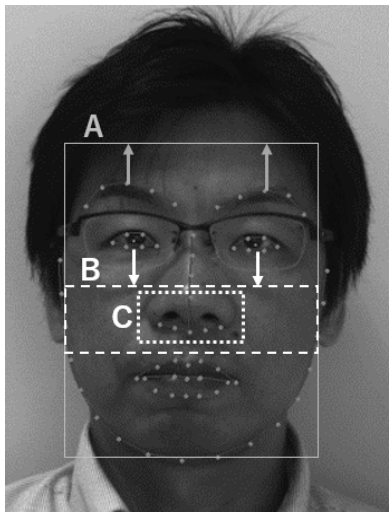


Fig. 2.3 Region of interest (A: Face, B: Cheek, C: Nose).

2.3.3 Acquisition of normalized time-series signals

Hemoglobin in blood has a strong property of absorbing the green wavelength of visible light (wavelength around 550 nm), and fluctuation of blood flow can be obtained from small temporal changes in green luminance value [21]. Therefore, in our method, the normalized intensity value of the green channel of each pixel is used to construct the time-series signal [22]. The normalization is done using equation (2.1).

$$g(t) = \frac{G(t)}{R(t)+G(t)+B(t)}, \tag{2.1}$$

where $g(t)$ is the green luminance value after normalization at time t, and $R(t)$, $G(t)$, and $B(t)$ are the red, green, and blue intensity values at time t, respectively.

2.3.4 Peak detection by first-order differential Gaussian function

The first-order differential Gaussian function $f(s)$ defined in equation (2.2) is convolved with the normalized time-series signal $g(t)$ using equation (2.3) to obtain the differential time-series signal $g'(t)$.

$$f(s) = \frac{s}{\sqrt{2\pi\sigma^3}} \exp\left(-\frac{s^2}{2\sigma^2}\right) \tag{2.2}$$

$$g'(t) = \sum_{s=-w}^{w} g(t+s) \cdot f(s) \tag{2.3}$$

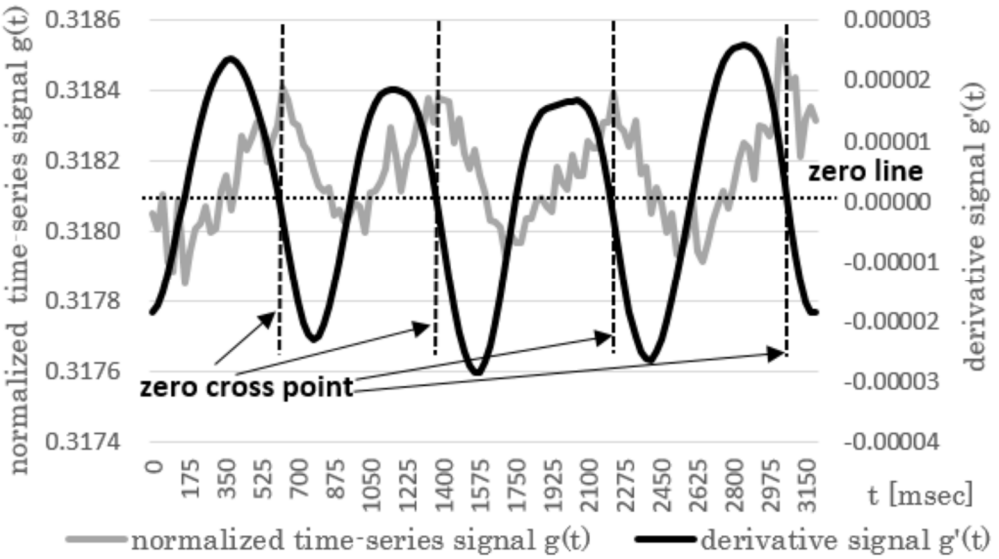


Fig. 2.4 Peaks detection method. Peaks are detected by finding zero-crossing points of the derivative of the normalized time-series signal.

Where, σ is the SD of the Gaussian function and w is the half of the window length of the convolution operation.

Next, the time at which the differential time-series signal $g'(t)$ switches from positive to negative (referred to as the zero-crossing point) is determined by equation (2.4), and is considered as the positive peak of the pulse wave. Fig. 2.4 shows an example of $g(t)$ and the differential time-series signal $g'(t)$.

$$g'(t-1) \geqq 0 \text{ and } g'(t) < 0 \tag{2.4}$$

2.3.5 Calculation of RRI by Multiple-Measurement-Points-Voting-Method

For the normalized time-series signal at each pixel in the ROI, let RRI be the time interval between two consecutive peaks, P_i and P_{i-1}. Here, we introduce an RRI map which represents the estimated RRI for each pixel. The RRI map has the same dimension as the face ROI. Fig. 2.5 shows an example of RRI calculation, and Fig. 2.6 shows an example of an RRI map.

Then, we find the most frequent RRI value in the RRI map as the representative RRI at the frame. If there is more than one value, the minimum value of RRI is used. The interval between consecutive RRI maps is the frame interval.

The above procedure is summarized below. Since the procedure is applied to each frame, real-time processing is possible. However, there is a time delay equal to the length of the convolution function. It is assumed that at least one peak has been detected in the normalized time-series signal of each pixel.

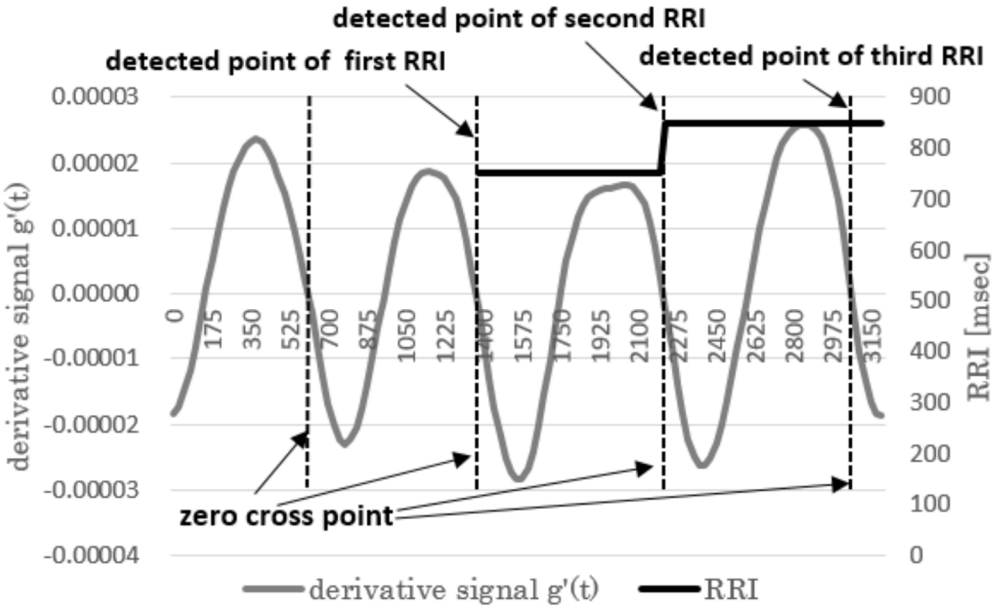


Fig. 2.5 RRI detection method. RRI is defined as the interval between the zero-crossing points.

800	800	1125	1100	800	800
950	725	800	975	800	850
825	700	800	850	875	950
800	750	800	775	725	1000
775	800	775	800	600	975
875	800	800	825	800	950
800	800	800	800	1200	1100
650	300	800	375	125	1025

Fig. 2.6 An example of RRI [msec] map. Assume that face ROI is W6×H8 pixels. Each pixel has the value of the estimated RRI at the pixel.

For each frame,

Step 1 Detect a peak by the first-order differential Gaussian function at each pixel.

Step 2 When a peak is detected, calculate the time difference (RRI) from the previous peak at the pixel, and update the value of the pixel in the RRI map.

Step 3 Calculate the distribution of RRI values in the RRI map, and find the most frequent value as the representative RRI at the frame.

Table 2.2 ROI size [pixel].

Data set		Face ROI	Cheek ROI
Stationary state	DS1	W:330~445 H:381~514	W:330~445 H:54~118
	DS2	W:322~460 H:384~500	W:322~460 H:52~127
Working state	DS3	W:237~355 H:313~427	—

2.4 Experimental Results

The spatial smoothing parameter, S, of the ROI was 61 pixel, the SD, σ, of the first-order differential gaussian function was 4, and the one-sided window width, w, of the convolution was 3σ. The position and size of the ROIs were determined experimentally. The upper edge of the face ROI (outer rectangle in Fig. 2.3) was set to 60 pixels above the eyebrows, and the left, right, and bottom edges were aligned with the contour points of the face. The upper edge of the cheek ROI (inner rectangle) was set to 60 pixels below the lower eyelids to exclude the region of the spectacles, the left and right edges were aligned with the facial contour points, and the bottom edge was 10 pixels above the top of the upper lip. The size of the ROI in the experimental data was determined according to the size of the subject's face and the distance from the camera, as shown in Table 2.2.

2.4.1 RRI estimation accuracy

The accuracy of RRI estimation of the proposed method was evaluated using the PPG signals that were acquired simultaneously with video images. The peak was detected from the PPG signals by calculating the local maxima. MATLAB's 'findpeaks' function was used to detect the peaks. Using the detected peaks, the true value of RRI for each frame was calculated as shown in Fig. 2.7. The heart rate, HR, in beats per minute (bpm) was calculated using equation (2.5).

$$HR = \frac{1}{\frac{1}{N-t_0+1}\sum_{f=t_0}^{N} RRI_f} \times 60, \tag{2.5}$$

where RRI_f is the RRI of the fth frame. The evaluation start frame t_0 was set to 121 frames, and N was set to 7200 (the total number of frames in 3 minutes of recorded video). As the RRI can't be calculated without the two consecutive heartbeats and the time for two consecutive heartbeats is 3 sec (120 frames), considering a minimum heartbeat of 40 bpm, t_0 is set to 121 frames.

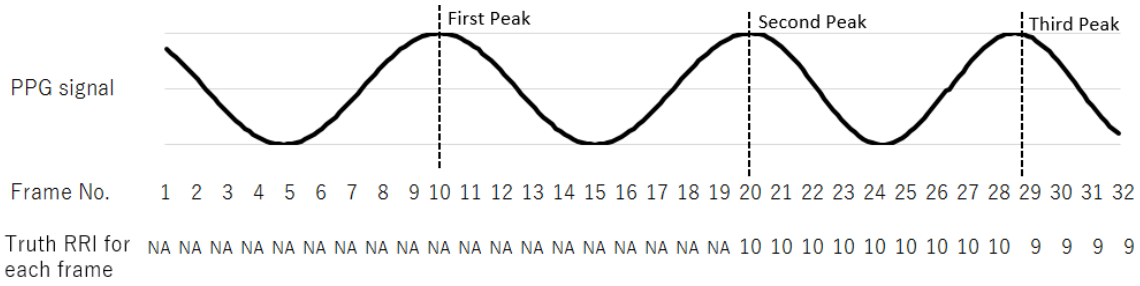


Fig. 2.7 Example of deriving truth RRI for each frame. NA means that available RRI is not calculated.

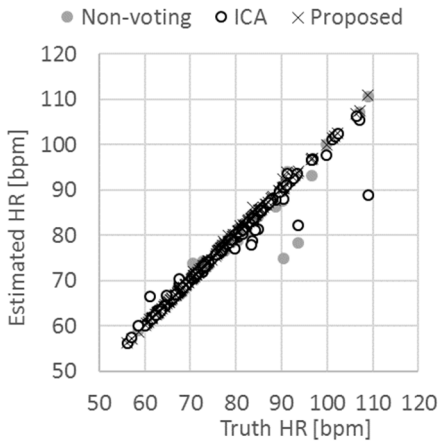
We compared our proposed method with two conventional methods. The first method was ICA based method (hereinafter referred to as the ICA method) [8]. The second method calculated RRI using the first-order differential Gaussian function on the time-series signal obtained by averaging the entire ROI (hereinafter referred to as the Non-voting method).

The scatter plots of the true HR and the estimated HR obtained by equation (2.5) are shown in Fig. 2.8, and their correlation coefficients are shown in Table 2.3. Next, the mean and SD of the mean absolute error (MAE) calculated by equation (2.6) and the p-value are shown in Fig. 2.9. Student's t-test was used for significance test.

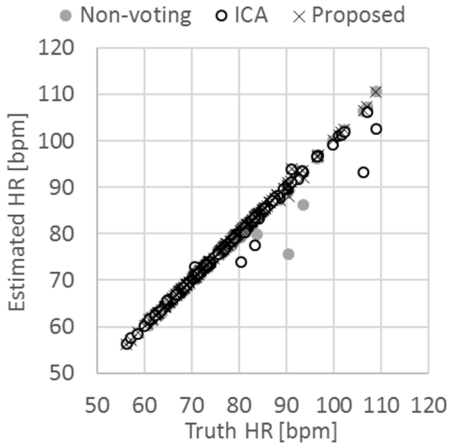
$$MAE = \frac{1}{N-t_0+1}\sum_{f=t_0}^{N}|r_f - p_f|, \tag{2.6}$$

where r_f and p_f are the true RRI and estimated RRI of the fth frame, respectively.

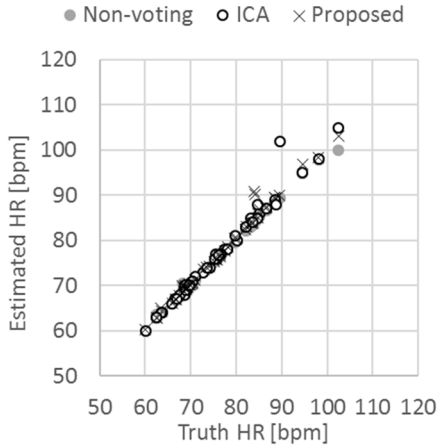
In terms of the accuracy of heart rate (Fig. 2.8 and Table 2.3), the proposed method and the conventional method were not significantly different. However, in terms of the MAE (Fig. 2.9), which compares the RRI measurement error for each frame, the proposed method showed significantly higher accuracy than the other conventional methods in all data sets and ROIs (significance level ≤ 0.05). The MAE of the proposed method was not significantly different among the data sets (significance level ≤ 0.05). Furthermore, the MAE (mean ± SD) of the proposed method for the face ROI, which was calculated by dividing DS2 into those with and without spectacles, was 22.5 ± 10.9 for those with spectacles and 21.3 ± 10.6 for those without spectacles, with no significant difference (significance level ≤ 0.05).



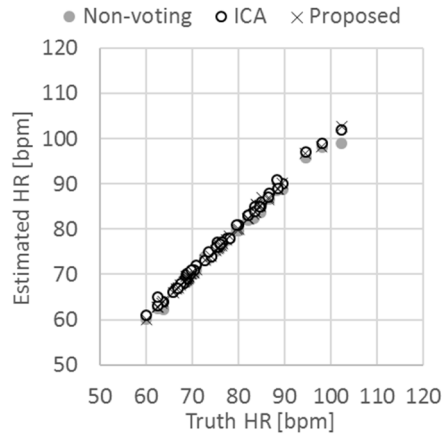
(a) DS1(Face ROI).



(b) DS1(Cheek ROI).



(c) DS2(Face ROI).

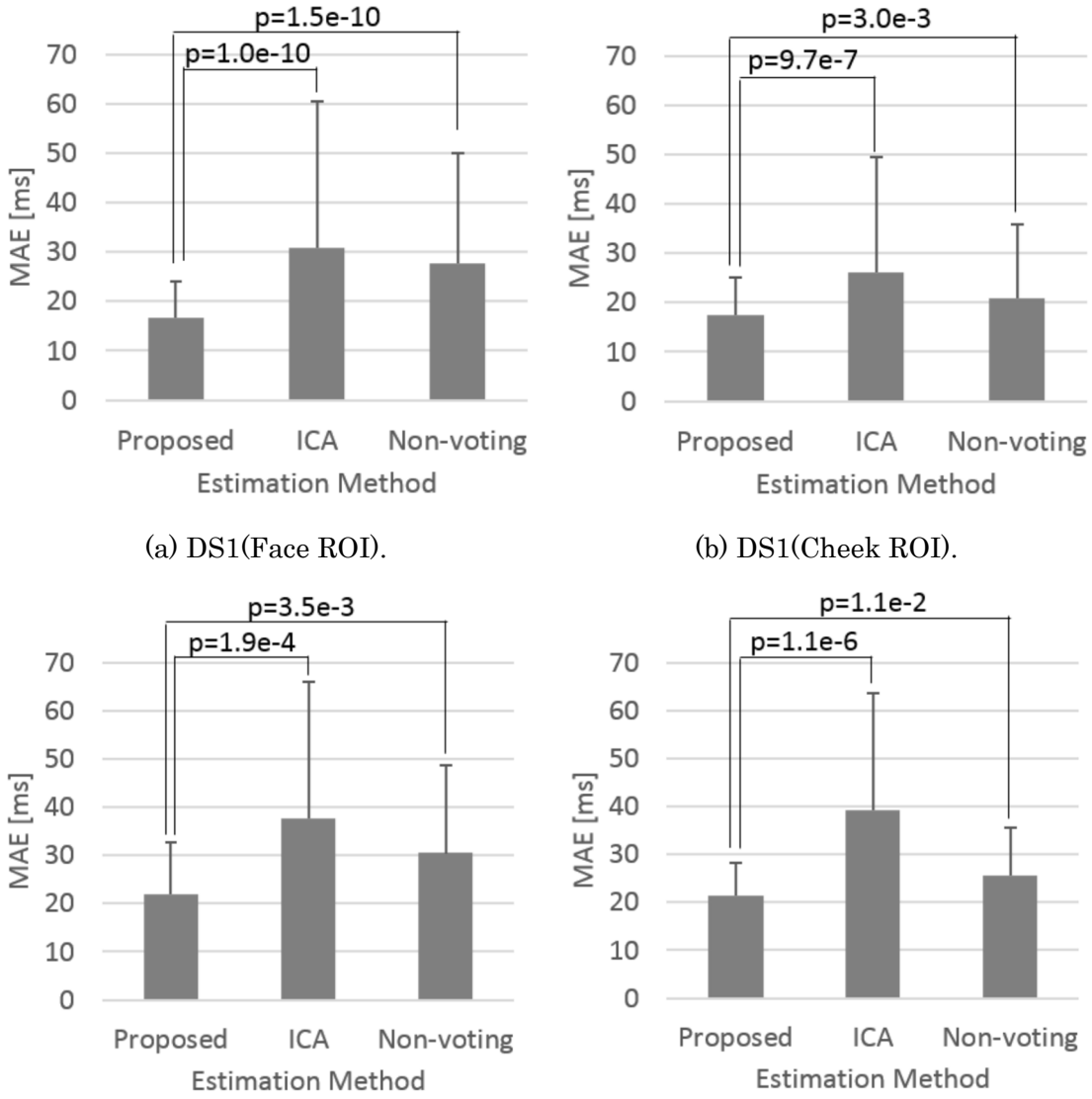


(d) DS2(Cheek ROI).

Fig. 2.8 Scatter plots between the estimated HR and the true HR using DS1 and DS2.

Table 2.3 Correlation coefficients between the estimated HR and the truth HR.

Data set	ROI	Estimation method		
		Proposed	ICA	Non-voting
DS1	Face	0.999	0.984	0.988
	Cheek	0.999	0.993	0.993
DS2	Face	0.992	0.987	0.997
	Cheek	0.999	0.998	0.998



(a) DS1(Face ROI). (b) DS1(Cheek ROI).

(c) DS2(Face ROI). (d) DS2(Cheek ROI).

Fig. 2.9 Mean and SD of MAE using DS1 and DS2.

Fig. 2.10 shows the mean, SD, and p-value of MAE for 60 videos from 10 subjects in DS3. Fig. 2.11 shows the MAE of each method for each of the 60 videos and the number of times the coordinate values of the nasal region moved by more than 3 pixels between frames. As shown in Fig. 2.10, the accuracy of the proposed method was significantly improved compared to the conventional methods even in the working state condition. Furthermore, the variation between the MAE data of the proposed method was as small as 24.7 msec. Fig. 2.11 shows that, the MAE of the proposed method was stable and had a high accuracy even for the data with many quick movements of more than 3 pixels per frame.

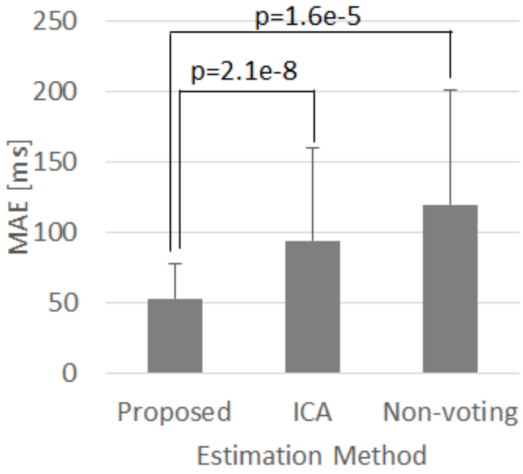


Fig. 2.10 Mean and SD of MAE in DS3.

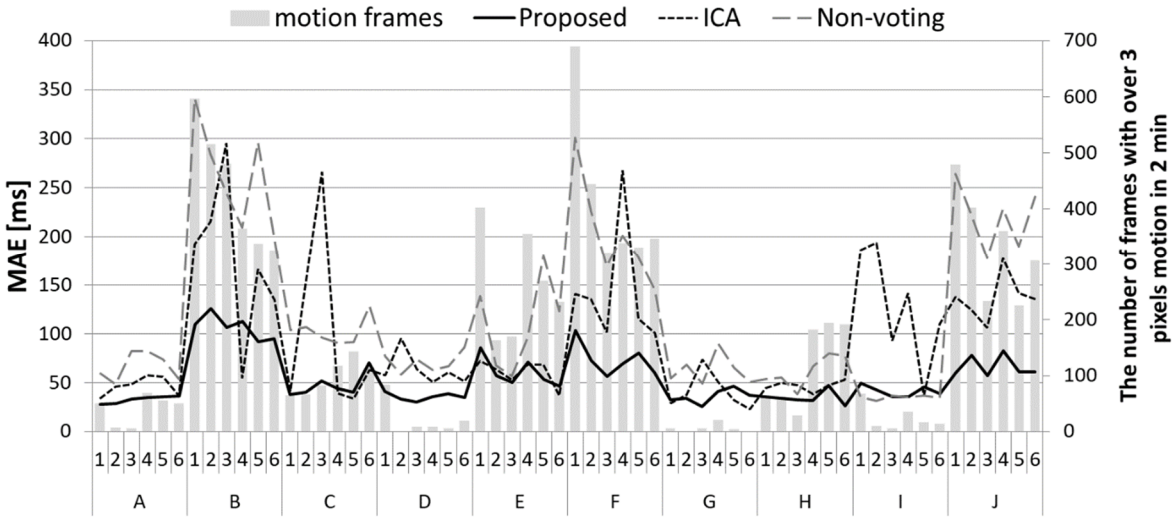


Fig. 2.111 Mean and SD of MAE in DS3. (A-J denote 10 different subjects, and numbers denote trial number).

2.4.2 Minimum measurement duration

In the previous section, we evaluated the estimation accuracy of the RRI except for the first three seconds. In this section, we compare and evaluate the ICA method and the proposed method for the time until satisfying the target accuracy using the stationary data shown in Section 2.1. Since the ICA method requires a minimum observation time of 2 sec, we started from 2 sec and evaluated at 0.5 sec intervals, as shown in Fig. 2.12. This figure shows the time-series signals separated by ICA at the time of each evaluation. It can be seen that as the analysis time becomes longer,

the time-series components corresponding to the change in PPG signal are obtained becomes better. The target accuracy was set to be within 50 msec of the error between the true RRI and the estimated RRI, and the accuracy was evaluated using the RRI at the end of the obtained results. The minimum target error between the true RRI and the estimated RRI was set to be within 50 msec as we considered the achievable accuracy to be double of the RRI measurement resolution of 25 msec for the data taken at 40 fps.

Fig. 2.13 shows the required time to satisfy the target accuracy for DS1 and DS2 in the face and cheek ROIs, respectively. In both ROIs for both data sets, the proposed method significantly reduces the time to reach the target accuracy compared to the ICA method (significance level ≤ 0.05).

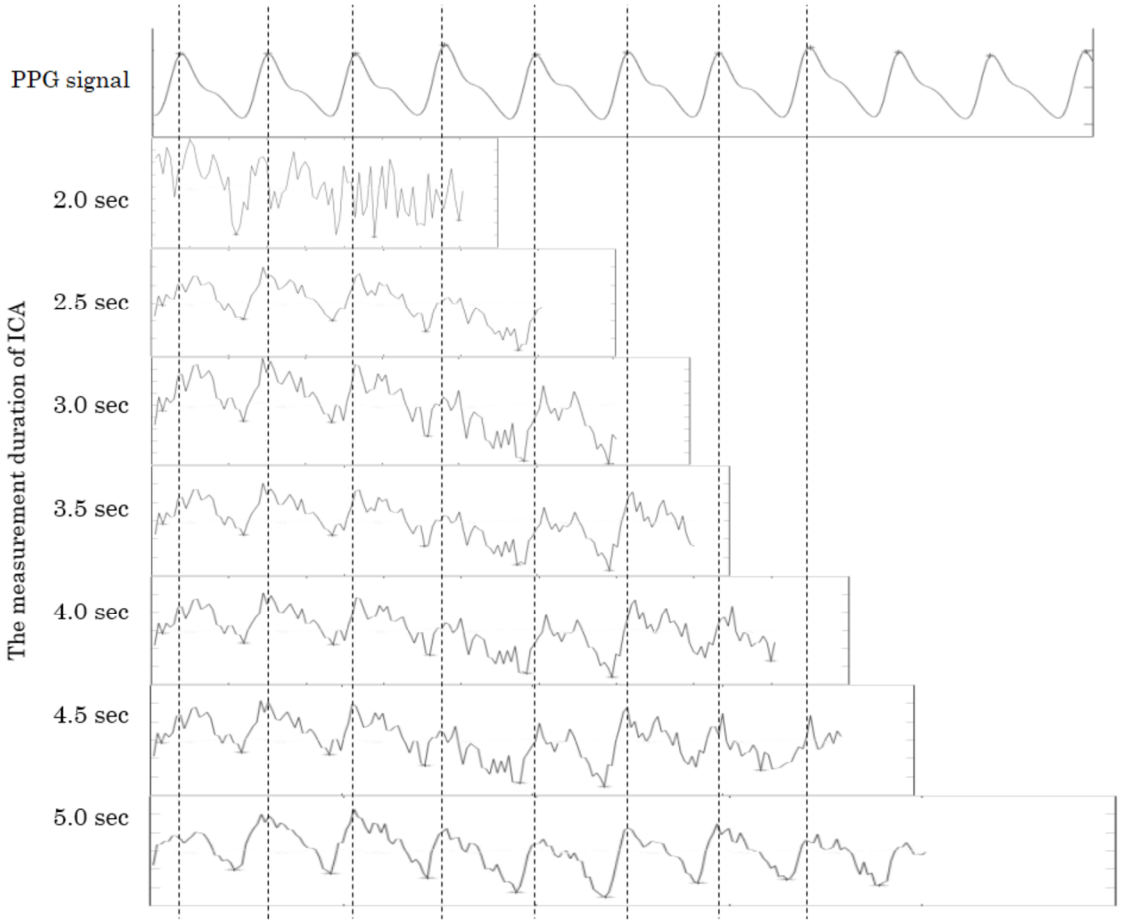
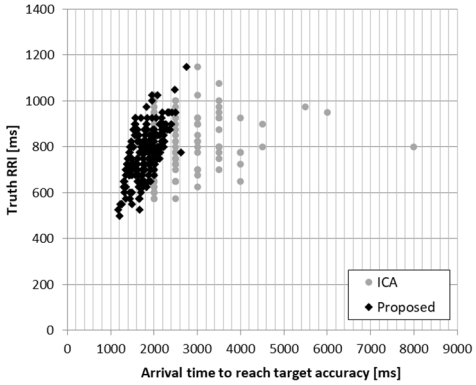
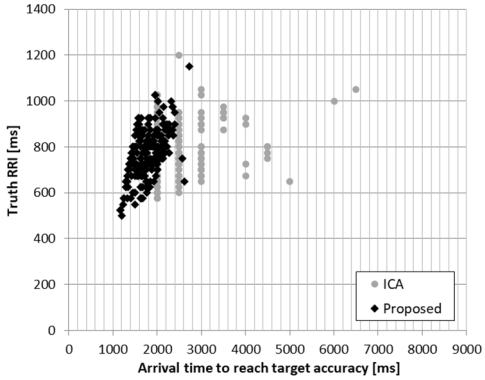


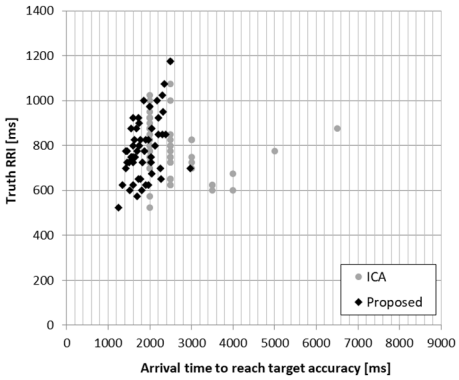
Fig. 2.12 Example of the ICA result for the different measurement duration



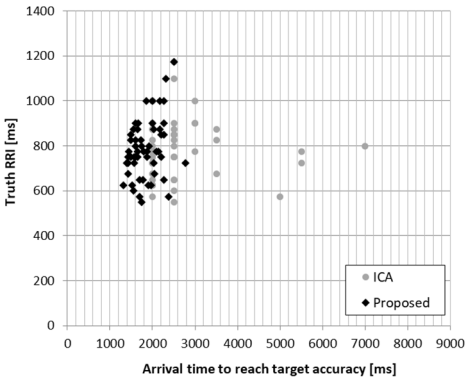
(a) DS1 (Face ROI).



(b) DS1 (Cheek ROI).



(c) DS2 (Face ROI).



(d) DS2 (Cheek ROI).

Fig. 2.132 Comparison of the minimum measurement duration to satisfy target accuracy between the proposed method and ICA method using DS1 and DS2.

2.4.3 Computation time

The average computation time per frame for 300 frames of each of the three sample data with different sizes of ROI is shown in Table 2.4. The average computation time was measured separately for the following three processes. The specification of the computer was Core i7-4790 3.6GHz memory 8GB.

Process 1 ROI extraction: Processing of Section 2.3.2

Process 2 RRI calculation of each pixel: Step 1 and 2 in Section 2.3.5

Process 3 Voting: Step 3 in Section 2.3.5

Table 2.4 Average processing time of 3 sample data.

Sample data		Average time of each process		
No	ROI size W×H [pixel]	ROI extraction	RRI calculation of each pixel	Voting
1	247×307	11.0 msec	14.5 msec	1.6 msec
2	302×384	11.7 msec	23.3 msec	6.1 msec
3	357×424	12.3 msec	31.7 msec	9.2 msec

2.5 Discussion

2.5.1 RRI estimation accuracy

Fig. 2.9 shows the results of the significant difference test between the proposed method and the conventional methods (ICA method and non-voting method) on the MAE results. The results show that the proposed method significantly reduced MAE more than the conventional methods in both DS1 and DS2, which were conducted with a small number of subjects and a large number of subjects, respectively. This indicates that the proposed method is robust both within and between individuals.

In both of DS1 and DS2, the MAE of the proposed method was significantly lower than that of the conventional methods in both of the cheek ROI (skin only) and the face ROI (including spectacles and hair in some people). There was no significant difference between the ROIs in either dataset, and the proposed method has the advantage that it does not require the process of limiting the ROI to the cheek ROI.

On the other hand, as shown in Table 2.3, the accuracies of heart rate estimation for all methods were high with a small difference between the proposed method and both of the conventional methods. However, the accuracy of frame-by-frame RRI estimation of the proposed method was significantly higher than the conventional method. This indicates that although the conventional method can obtain the estimated pulse rate with high accuracy, the deviation of the peak time from the true value is large. Since sympathetic and parasympathetic nerve balance is evaluated by frequency spectrum analysis of RRI, the proposed method is particularly effective in the analysis requiring RRI.

2.5.2 Minimum measurement duration

Table 2.5 shows the comparison of the mean and SD of the minimum measurement

time between the proposed method and the ICA method. The student's t-test was performed. Table 2.5 shows that the proposed method significantly reduced the minimum measurement time compared to the ICA method for all data sets and ROIs.

Table 2.5 Comparison of the minimum measurement duration to satisfy target accuracy between the proposed method and ICA method (mean±SD) [msec].

Data set	ROI	Estimation method		p-value between methods
		Proposed	ICA	
DS1	Face	1785±314	2419±766	4.1e-22
	Cheek	1781±309	2396±710	5.3e-24
DS2	Face	1856±340	2704±1365	2.3e-5
	Cheek	1843±329	2583±1018	1.6e-6

Next, to estimate the observation time required for reliable RRI estimation, we define the required observation time as the time required for 3σ data to achieve the target accuracy. This is the time when 99.7% of the experimental data satisfy the target accuracy, that is, the shortest measurement time. Table 2.6 shows the required observation time obtained using a total of 252 trials of DS1 and DS2 data. The results of the face ROI in Table 2.6 shows that the required observation time is about 2.8 sec for the proposed method, and one is about 5.3 sec for the ICA method. So, the proposed method can be applied with half of the input signal length of the window width of the first-order differential Gaussian function in addition to the minimum two beats required for RRI estimation, while the ICA method requires sufficient input signal length.

Table 2.6 Minimum measurement duration to reach target accuracy [msec].

	Proposed	ICA
Face	2,762	5,286
Cheek	2,737	4,802

In order to detect pulse waves using face images, it is necessary to extract images of the same part of the body for a certain period of time from a moving object such as a person's face. For example, when images are taken by a camera on a computer monitor, the face always faces the front, so it is easy to extract the same part of the face for a certain period of time. When image is taken by a surveillance camera of a

person in daily life, problems occur such as the angle of the face changes. Even in daily life, people do not move all the time and are often in a stationary state. Therefore, a method with a short observation time is desirable to estimate the RRI in the same static state. The proposed method can estimate RRI in about 47% less observation time than the ICA method, which means that the proposed method is highly practical.

2.5.3 Calculation time

Each of the processes in Table 2.4 (ROI extraction, RRI calculation of each pixel, and Voting) can be done within 33 msec, which is the frame interval of a 30 fps camera. Therefore, we developed a software that processes each method in parallel on multiple cores. The real-time operation was confirmed for the data in Table 2.4, No.1 and No.2, which were captured using a 30 fps camera. Each process can be done independently, but the results of the first process are used in the next process. Therefore, the throughput required for real-time processing was achieved, but latency was incurred.

2.6 Summary

In this chapter, we have shown that the multipoint measurement voting method, which estimates the RRI from the most frequent values of the RRI in the region, was significantly more accurate than the conventional ICA and non-voting methods. In addition, the observation time required for reliable estimation of RRI was reduced by about 47% compared to the ICA method. Furthermore, with parallel computation, we achieved real-time processing for data with ROI size within 302×384 pixels.

We also evaluated the proposed method for persons in working state. It showed that the proposed method is significantly more accurate than the conventional method on the persons with motion. However, the accuracy in the PC working state was significantly lower than that in the stationary state. In future, we will investigate the effect of motion on the distribution of RRI values and develop a method that is more robust against motion. We will also investigate the accuracy of RRI estimation necessary for stress calculation, and aim to develop a method that can calculate stress with high accuracy in practical scenes.

Chapter 3 : Facial skin blood perfusion change based liveness detection using video images

3.1 Introduction

Recently, facial images have been used widely as a user-friendly person authentication method. However, spoofing has become a problem. Main spoofing techniques include printed images (photographs), still and video images on flat monitors and screens, and 3D masks and printed 3D faces similar to a person's faces. Conventional anti-spoofing methods use facial recognition cameras capturing visible wavelength or other devices such as IR cameras and range camera.

Spoofing detection methods using only facial recognition cameras can be divided into two categories – methods that deal with spoofing using single still images, and methods that address spoofing using video images. Regarding techniques dealing with spoofing using still images, there are some methods based on texture analysis using Local Binary Pattern (LBP) [23] and multiscale directional features using the shearlet transform [24]. The former showed a recognition rate (the rate of discrimination between biological and artificial objects) of 98.0% using printed images, and the latter showed a recognition rate of 88.9% using printed images or images shown in monitors. The common limitation of both methods was that 3D masks were not considered. Regarding techniques dealing with video images, a few methods were proposed for the detection of flat objects using optical flow [25] or focused images [26]. Both methods showed a recognition rate of 100% for printed images. In addition, a recognition rate of 91.7% was achieved for printed images with a method combining image texture, face background, and oscillating components [27]. Another method using Weber Local Descriptor (WLD) was proposed to deal with spoofing with both still and video images that achieved a recognition rate of 92.3% [28]. However, no evaluation results were mentioned for spoofing with 3D masks. To deal with the spoofing using 3D masks, liveness detection methods were proposed based on blinking [29]. The limitation of these methods is that specific actions are required for these methods to work, which would impair the user's convenience.

Using other images such as range images [30], thermal images [31], near-infrared (IR) images [32], some anti-spoofing techniques have been proposed. The recognition

rate of a method based on facial curvature distribution using range images was 100% for inflected printed images [30]. A method based on thermal images detects the spoofing by calculating a correlation with segmented visible facial images [31]. The false acceptance rate was 0.1% and recognition accuracy was 85.1% (90.8% for faces without spectacles). This method, however, lacks practicability as a special camera is required. Regarding a method using near IR images [32], texture analysis was applied to near IR images and visible images, thus providing tolerance for illumination changes. A recognition rate of 99.8% was achieved for printed images and still monitor images, but evaluation using 3D masks was not performed.

Recent studies explore the detection of pulse waves from color images. Takano *et al.* [17] analyzed frequency spectra of intensity variation in the cheek area to confirm coincidence with pulse periods measured by a pulse oximeter. Poh *et al.* [10] applied ICA and frequency spectrum analysis to RGB signals in the facial area to improve calculation accuracy of pulse period. They also reported that green channel of the RGB images offers the highest sensitivity to pulsation. These studies are interesting as they describe ways of detecting changes in blood perfusion from facial images. However, multiple pulse waves (e.g., several periods of ten seconds) are used for heart rate calculation, which is not suitable for liveness detection in facial authentication systems aiming at fast recognition.

This chapter proposes a spoofing detection method that addresses 3D objects by detecting the blood perfusion changes within a short time from video images acquired by facial recognition cameras. The proposed method rapidly detects biological response through evaluation of intensity-based time series in terms of likeness to biological pulse waves. Performance of the method is evaluated using 16 real (human) subjects, 5 illumination conditions, and 4 types of artifacts including 3D objects, still images, and video images.

3.2 Experimental Data

This study was conducted with the approval of Ethics Committee of University of Hyogo as a “Liveness detection using facial images of video camera”.

3.2.1 Data acquisition equipment

Ten-second videos of real subjects and spoofing objects were captured using a color camera (Basler acA1920-155uc, sensor: Sony IMX174, resolution: 1920 × 1200, 30 fps). The captured images were recorded as non-compressed AVI files.

3.2.2 Real subjects

Videos were acquired three times for a total of 16 healthy males and females, two in each age group of 20s, 30s, 40s, and 50s, under five illumination conditions (L1-L5) described below. The subjects stood in front of the fixed camera at a certain position (about 50 cm away from the camera), and looked at the camera for a certain time period. Shooting started before a subject stood in front of the camera, and lasted for several seconds. It took for a subject 2-4 sec to stand in front of the camera, and the total shooting time was 10 sec. Shooting under different illumination conditions is illustrated in Fig. 3.1. The lighting was normal white light.

(L1) Room light (without the presence of sunlight): 380 lux

(L2) Sunlight in front of face (no other light): [2,130, 6,400] lux

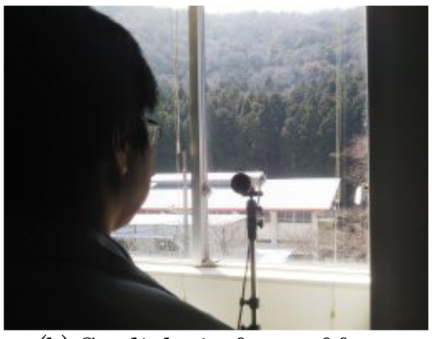
(L3) Sunlight on back of head (no other light): [140, 420] lux

(L4) Sunlight on back of head with room light: [530, 740] lux

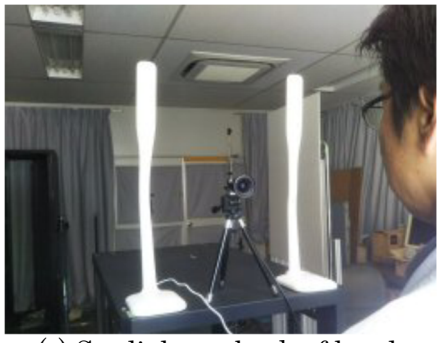
(L5) Sunlight on back of the head with room light and facial light: [1080, 1360] lux

The above illumination values are the minimum and the maximum recorded values during the video acquisition. The values were measured by a lux meter placed at 50 cm from the camera.

(a) Room light.



(b) Sunlight in front of face.



(c) Sunlight on back of head.

(d) Sunlight on back of head with room light.

(e) Sunlight on back of head with room light and facial light.

Fig. 3.1 Experimental environments.

3.2.3 Artificial subjects

The lighting environment was the same as in L1 (room light only, 380 lx). The recording started before an operator took a spoofing object and approached the camera, and lasted for a certain period while the operator was standing in front of the fixed camera at a certain position (so that the operator's face was the same size as the spoofing object); the total recording time was 10 sec. The experimental setup is demonstrated in Fig. 3.2.

The spoofing objects held in hand to create video data of artificial subjects are described below (O1-O4). 12 data were obtained for O1, and 3 data were obtained for

each of the O2, O3, and O4. The same experiments were conducted by 4 different operators, and a total of 768 data were obtained.

(O1) Four types of dolls (Fig 3.3).

(O2) Printed images of 16 persons.

(O3) Still images of 16 persons displayed on an LCD monitor.

(O4) Video images o 16 persons displayed on an LCD monitor.

The LCD monitor was 13.1 inches wide, with resolution of 1600 × 900 and refresh rate of 59 Hz. The spoofing objects, O2-O4, were images of the subjects described in the previous subsection.

(a) Doll.

(b) Printed image.

(c) Monitor image.

(d) Monitor video.

Fig. 3.2 3 3-D (a) and 2-D (b-d) spoofing data collection.

Fig. 3.3 Four kinds of doll.

3.3 Proposed Method

3.3.1 Overview

In a captured video, a frame where the user's face is facing front is selected as the first frame to be analyzed. Three time-series signals of green, and blue (RGB) color channels in the face ROI are extracted. The time-series signals are normalized (referred to as normalized time-series signals) by dividing them with the sum of three time-series signals. Next, a the first-order derivative Gaussian function is convolved. The obtained signal is referred to as normalized time-series derivative signals. The method then extracts two features for liveness identification from the normalized time-series derivative signals. The first feature is the correlation coefficient between the red and the green normalized time-series derivative signals (Referred to as R-G correlation feature). The second feature is the correlation coefficient between the green normalized time-series derivative signals between two regions (Referred to as inter-area correlation feature). Finally, spoofing is detected by using support vector machine (SVM) with the extracted features. The flow chart of the proposed method is shown in Fig. 3.4.

3.3.2 Region of interest extraction

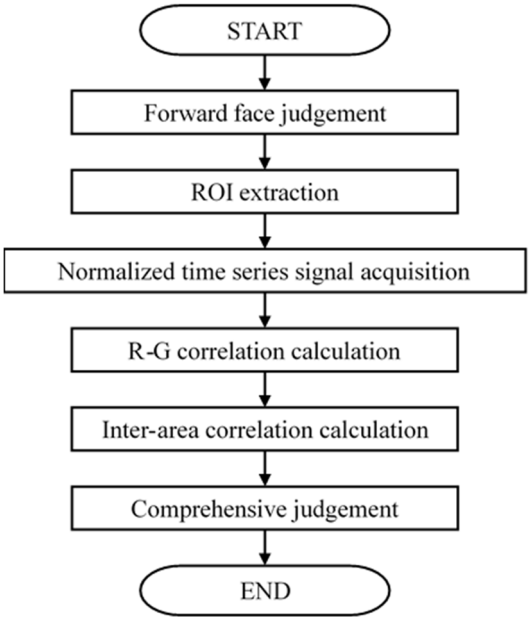


Fig. 3.4 4 Flow chart of the proposed method.

Facial feature points (shown by dots in Fig. 3.5) are detected using Dlib C++

library [19]. In total, 68 feature points are detected. Using coordinates of the detected feature points, frontal face is recognized when the following condition met.

$$\left|\frac{(Lx+Rx)}{2} - Nx\right| < \frac{(Rx-Lx+1)}{10}, \tag{3.1}$$

where Nx is x coordinate of the nose center, Lx is x coordinate of the leftmost face contour, and Rx is x coordinate of the rightmost face contour.

When the frontal face is recognized, Total of ROI (ROI T) is defined by equations (3.2)-(3.5). I_W, I_H, I_x, I_y are width and height of ROI T, x, and y coordinates of the upper left vertex, respectively.

$$I_W = \left\{\frac{(Rx - Lx + 1)}{20}\right\} \times 18 \tag{3.2}$$

$$I_H = \left\{\frac{(Rx - Lx + 1)}{20}\right\} \times 3 \tag{3.3}$$

$$I_x = \left\{\frac{(Rx - Lx + 1)}{20}\right\} + Lx \tag{3.4}$$

$$I_y = \left\{\frac{(My - Ny + 1)}{2}\right\} - I_H \tag{3.5}$$

Where, My is y coordinate of the lip center, and Ny is y coordinate of the nose center. Besides, ROI T is equally divided into subregions A, B, C in the horizontal direction, from left to right.

The subregion B is used as a template image to track the object's movement by correcting ROI position in each frame using the template matching. Correlation

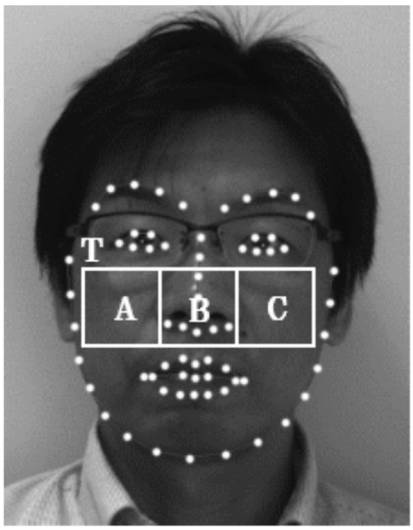


Fig. 3.5 5 Region of interests. T is the total area of A, B, and C.

coefficient is used as a likelihood function. Besides, when the detected movement exceeds a certain value from the previous frame (k pixels in x or y direction), the frontal face detection algorithm is applied again for that frame. Here k is an empirical analysis parameter.

3.3.3 Normalized time-series derivative signals

For each region T and subregions A, B, and C, normalized time-series derivative signals are derived as follows. In each region, mean time-series signals ($R(t)$, $G(t)$, $B(t)$) for each channel are obtained. Then, color component ratio, which is called normalized time-series signals, $g(t)$ and $r(t)$, are calculated using equations (3.6) and (3.7) respectively.

$$g(t) = \frac{G(t)}{R(t) + G(t) + B(t)} \tag{3.6}$$

$$r(t) = \frac{R(t)}{R(t) + G(t) + B(t)} \tag{3.7}$$

Next, a first-order derivative Gaussian function, $f(s)$, defined by equation (3.8), is convolved to $g(t)$ and $r(t)$. $g'(t)$ and $r'(t)$ are normalized time-series derivative signals.

$$f(s) = \frac{s}{\sqrt{2\pi}\sigma^3} \exp\left(-\frac{s^2}{2\sigma^2}\right) \tag{3.8}$$

$$g'(t) = \sum_{s=-w}^{w} g(t+s) \cdot f(s) \tag{3.9}$$

$$r'(t) = \sum_{s=-w}^{w} r(t+s) \cdot f(s) \tag{3.10}$$

Here σ is the SD of the Gaussian function, and w is half-length of the window. In the same manner, normalized time-series derivative signals of subregions A, B, and C are obtained.

3.3.4 R-G correlation feature

Erythrocytes in blood flow show strong absorption of green light (around 550 nm) and weak absorption of red light (around 650 nm) [21]. As a result, color components in the facial skin change with an increase or decrease in blood flow. Therefore, R-G correlation feature, C_{RG}, that is the correlation coefficient between normalized time-series derivative signals of red and green intensity values is defined as a feature for

liveness detection.

$$C_{RG} = \frac{\sum_{t=1}^{n}\left(g_T'(t)-\overline{g_T'}\right)\left(r_T'(t)-\overline{r_T'}\right)}{\sqrt{\sum_{t=1}^{n}\left(g_T'(t)-\overline{g_T'}\right)^2}\sqrt{\sum_{t=1}^{n}\left(r_T'(t)-\overline{r_T'}\right)^2}}, \tag{3.11}$$

where $g_T'(t)$ and $r_T'(t)$ are normalized time-series derivative signals at ROI T, $\overline{g_T'}$ and $\overline{r_T'}$ are mean values, and n is the number of samples of the normalized time-series signals.

3.3.5 Inter-area correlation feature

Blood flow of the human simultaneously changes in any part of ROI. That is, the changes of intensity values due to the light absorption characteristics of blood are synchronized among multiple areas. Intensity change caused by blood flow is the most pronounced for green color; hence we used intensity values of the green channel. So, inter-area correlation coefficient, C_{area}, defined in equations (3.12)-(3.14), will be effective to detect liveness.

$$C_{AB} = \frac{\sum_{t=1}^{n}\left(g_A'(t)-\overline{g_A'}\right)\left(g_B'(t)-\overline{g_B'}\right)}{\sqrt{\sum_{t=1}^{n}\left(g_A'(t)-\overline{g_A'}\right)^2}\sqrt{\sum_{t=1}^{n}\left(g_B'(t)-\overline{g_B'}\right)^2}} \tag{3.12}$$

$$C_{BC} = \frac{\sum_{t=1}^{n}\left(g_B'(t)-\overline{g_B'}\right)\left(g_C'(t)-\overline{g_C'}\right)}{\sqrt{\sum_{t=1}^{n}\left(g_B'(t)-\overline{g_B'}\right)^2}\sqrt{\sum_{t=1}^{n}\left(g_C'(t)-\overline{g_C'}\right)^2}} \tag{3.13}$$

$$C_{AC} = \frac{\sum_{t=1}^{n}\left(g_A'(t)-\overline{g_A'}\right)\left(g_C'(t)-\overline{g_C'}\right)}{\sqrt{\sum_{t=1}^{n}\left(g_A'(t)-\overline{g_A'}\right)^2}\sqrt{\sum_{t=1}^{n}\left(g_C'(t)-\overline{g_C'}\right)^2}} \tag{3.14}$$

Where, $g_A'(t)$, $g_B'(t)$, and $g_C'(t)$ are normalized time-series derivative signals for the subregions A, B, and C. We introduce a new feature which is defined as an average of correlation coefficients, C_{AB}, C_{BC}, C_{AC} as given by equation (3.15).

$$C_{area} = \frac{C_{AB} + C_{BC} + C_{AC}}{3} \tag{3.15}$$

3.3.6 Spoofing detection with pattern recognition

R-G correlation and inter-area correlation are obtained as image features expressing liveness. Using these features, spoofing is detected by solving a two-class discrimination problem with SVM, which is one of pattern recognition methods. Here

living objects are defined as class 1, and spoofed objects are defined as class 0. Pre-collected data are used as the training data for SVM. For the evaluation data, spoofing is detected if the output value of trained SVM is 0 or less.

3.4 Experimental Results

The method proposed in Section 3.2 was applied to all data described in Section 3.3. As regards parameters, the first-order derivative Gaussian function, SD, σ, was 2.4 frames, and the window length, w, of the first-order derivative Gaussian function was 5 frames. The movement threshold at template matching, k, was 5 pixels, and the evaluation time were 1 sec, 1.5 sec, 2 sec. The fps of the camera was 30. The evaluation time lengths, n, were 30, 45, 60 frames. ROI height (mean ± SD), IH, was 95.8±9.8 pixels for real subjects, 90.8±5.0 pixels for dolls, 106.7±13.3 pixels for printed images, 99.7±13.6 pixels for still monitor images, and 92.8±13.1 pixels for videos.

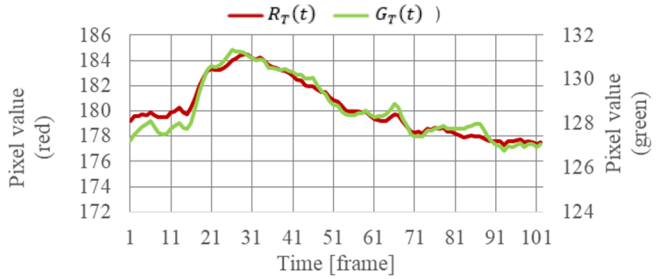
3.4.1 Time-series signals in region of interest

Fig. 3.6 shows an example of the time-series signals of average intensity, $R_T(t)$, $G_T(t)$ in the region T, the corresponding normalized signals, $r_T(t)$, $g_T(t)$, the normalized derivative signals $r_T'(t)$, $g_T'(t)$, and the normalized derivative signals $g_A'(t)$, $g_B'(t)$, $g_C'(t)$, in regions A, B, C for a real subject. Similarly, Fig. 3.7 shows an example of the normalized time-series signals of a doll, a printed image, a still monitor image, and a video. The normalized time-series derivative signals are plotted from the 11th frame because of a calculation delay corresponding to the window width (frames) in the convolution of the first-order derivative Gaussian function.

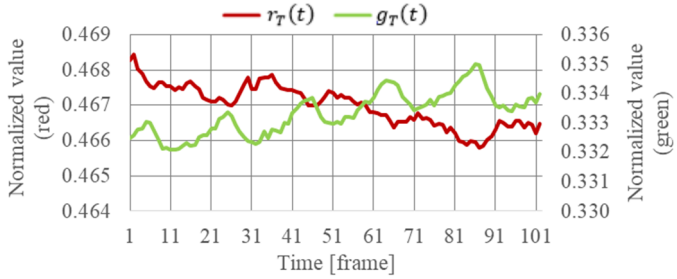
As shown in Fig. 3.6(b), in the case of real face, there is a negative correlation between $r_T(t)$ and $g_T(t)$. Moreover, Fig. 3.6(c) indicates that the negative correlation between $r_T'(t)$ and $g_T'(t)$ can be enhanced by convolution of first-order derivative Gaussian function defined by equations (3.8)-(3.10). In addition, positive correlations between all possible combinations of the normalized derivative signals ($g_A'(t)$, $g_B'(t)$ and $g_C'(t)$) were confirmed.

In contrast, as can be seen from Fig. 3.7, in the case of spoofed objects, there was neither a negative correlation between normalized time-series derivative signals in the whole region of interest or positive correlation between the subregions in

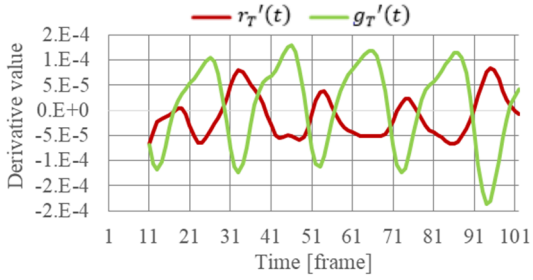
comparison with other than video images of doll, printed image, still monitor image.



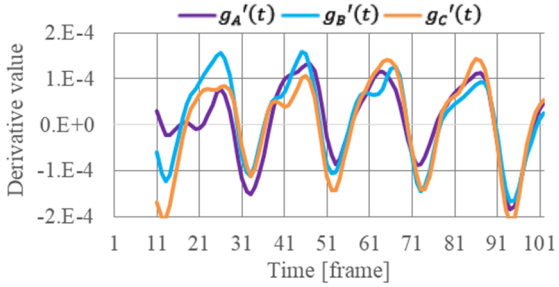
(a) Time-series signal $R_T(t)$ and $G_T(t)$.



(b) Normalized time-series signal $r_T(t)$ and $g_T(t)$.

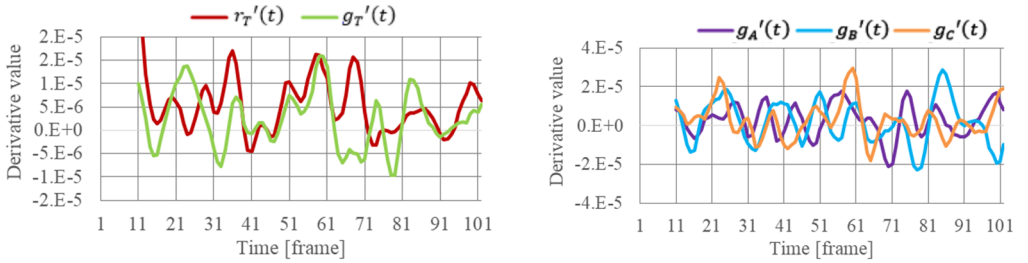


(c) Derivative signal $r_T'(t)$ and $g_T'(t)$.

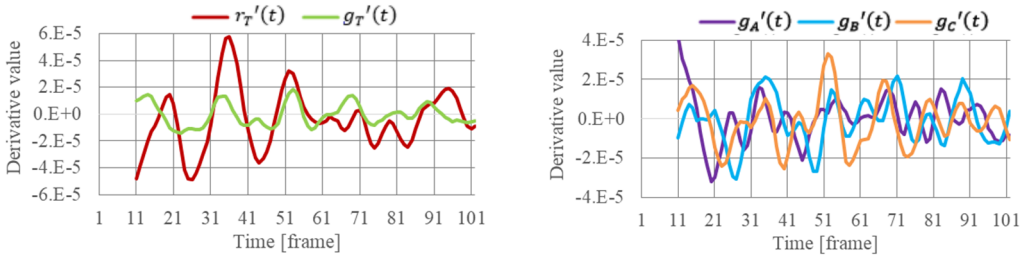


(d) Derivative signal $g_A'(t)$, $g_B'(t)$, and $g_C'(t)$.

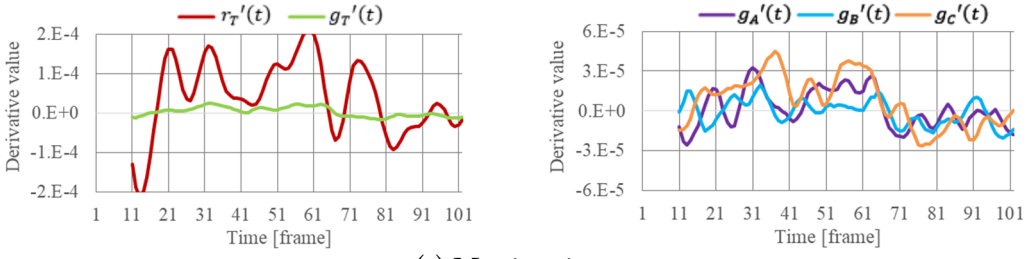
Fig. 3.6 Example of the signal of real subject.



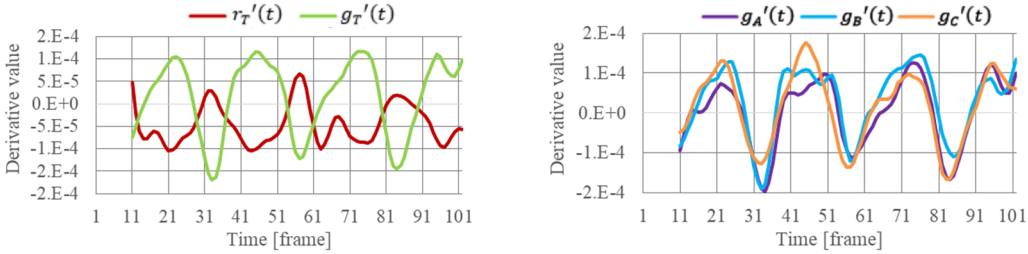
(a) Doll.



(b) Printed image.



(c) Monitor image.

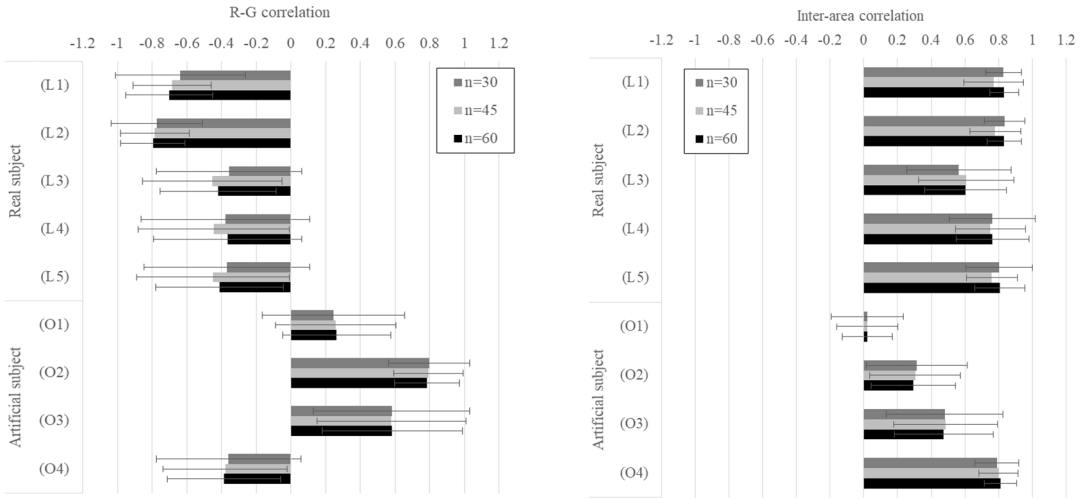


(d) Monitor video.

Fig. 3.7 Example of the signal of artificial subjects. Left figures show the derivative signal $\boldsymbol{r_T'(t)}$ and $\boldsymbol{g_T'(t)}$. Right figures show the derivative signal $\boldsymbol{g_A'(t)}$, $\boldsymbol{g_B'(t)}$, and $\boldsymbol{g_C'(t)}$.

3.4.2 Extracted features

Fig. 3.8 compares two features, R-G correlation and inter-area correlation, in the different illumination conditions and of the different spoofing objects. Here the evaluation time length, *n*, was varied in three periods—30 frames (1.0 sec), 45 frames (1.5 sec), and 60 frames (2.0 sec). Fig. 3.8(a) shows that the average R-G correlation was always negative for real subjects, and positive for spoofing objects, except for video (O4). A significant difference (significant level ≤ 0.01) was confirmed for the all combination of between real subjects (L1) and each spoofing objects (O1-O4). Fig.

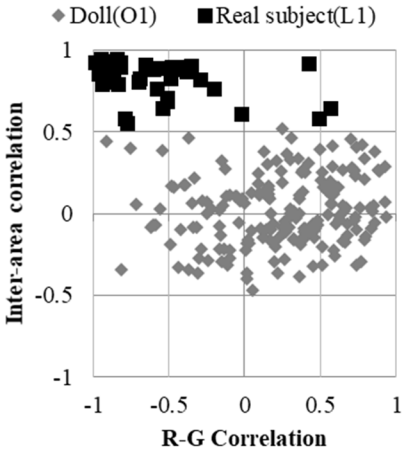


(a) R-G correlation, C_{RG}. (b) Inter-area correlation, C_{area}.

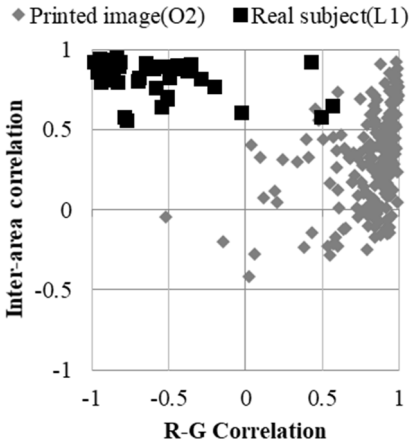
Fig. 3.8 Comparison of feature values. Vertical axis represents five kinds of lighting conditions (L1-L5), and four spoofing methods (O1-O4).

3.8(b) indicates that inter-area correlation was neither negative nor positive for doll (O1), and positive in all other cases. Particularly, this positive correlation was strong for real subjects and video (O4), while the SD was small. In addition, SD tended to decrease with a longer evaluation time. Significant difference (level: 0.01) was confirmed for the all combination of between real subjects (L1) and each spoofing objects (O1-O3).

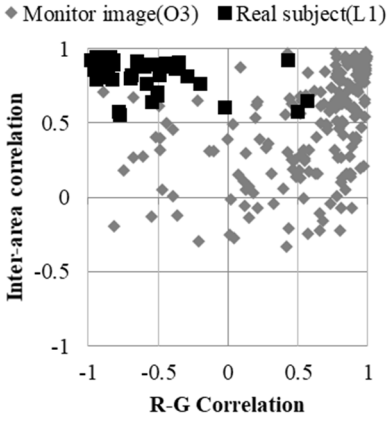
Fig. 3.9 represents plots of the two features for real subjects (3 data for each of the 16 subjects) and for spoofing objects (a total of 192 data for dolls, printed images, still monitor images, and videos). The horizontal axis shows R-G correlation feature, and the vertical axis does the inter-area correlation feature. Here the evaluation time length, n, was 30 frames. It can be seen from the plots that each spoofing object can be separated using either R-G correlation or inter-area correlation. However, both correlation features are required to distinguish spoofing objects from real subjects.



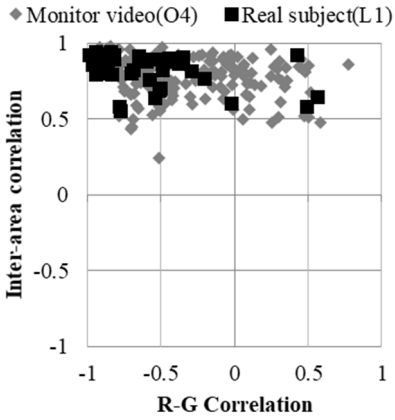
(a) Real subject and doll.



(b) Real subject and printed image.



(c) Real subject and monitor image.



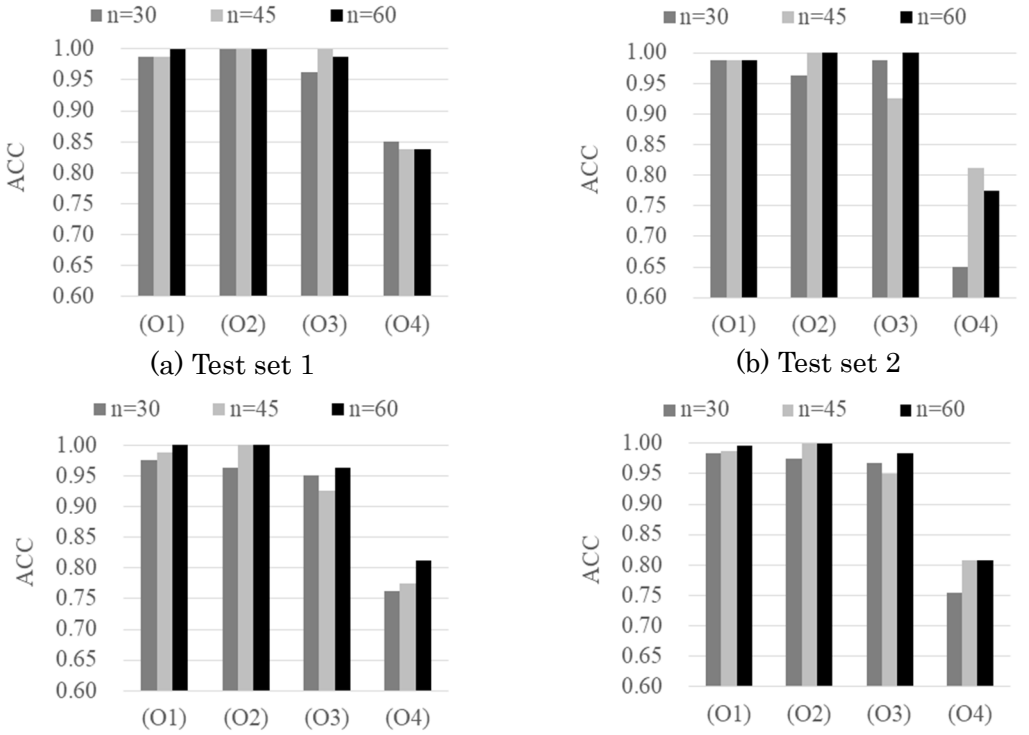
(d) Real subject and monitor video.

Fig. 3.9 Relationship between R-G correlation and inter-area correlation.

3.4.3 Liveness detection results

The results of spoofing detection obtained by SVM using R-G correlation and inter-area correlation features are described below. Evaluation was performed on real subjects (16 subjects, three data each), and 192 data of 3D spoofing objects (dolls) and plane spoofing objects (printed images, still images, and images on a monitor). The video data for both real subjects and spoofing objects were collected 3 times under each illumination condition. Among them, one video was used as evaluation data and the other two videos were used as learning data to conduct three-fold cross-validation. Recognition rates for the evaluation data are shown in Fig. 3.10. The recognition rate (ACC: accuracy) is defined in the following way.

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3.16}$$



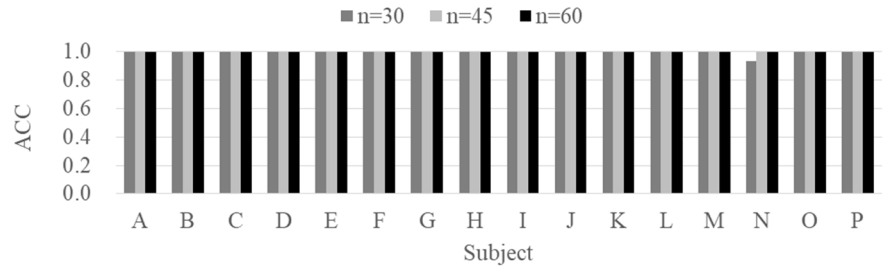
(a) Test set 1 (b) Test set 2

(c) Test set 3. (d) Average of three test sets.

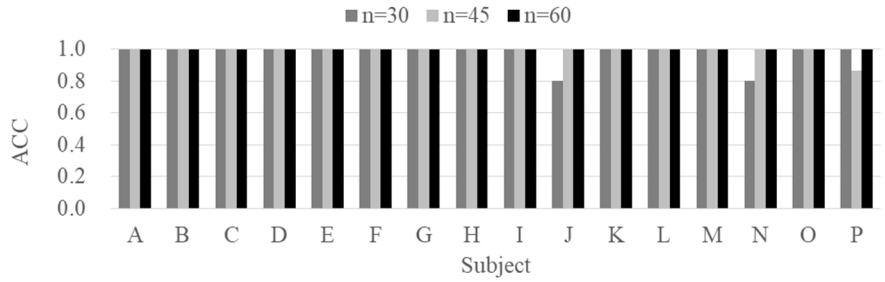
Fig. 3.10 Comparison of accuracy among four different spoofing attacks (O1-O4).

where TP is the number of correctly recognized real subject data (real subjects are recognized as real subjects), TN is the number of correctly recognized spoofing object (spoofing objects are recognized as spoofing objects), FP is the number of misclassified spoofing data (spoofing objects are recognized as real subjects), and FN is the number of misclassified real subject data (real subjects are recognized as spoofing objects). As indicated in the figures, all evaluation data, except for video (O4), were recognized at an accuracy of 90% or higher. With the evaluation time length n = 60 frames shown in Fig. 3.10(d), the recognition rate was 99.6% for dolls (O1), 100% for printed images (O2), and 98.3% for still images on monitors.

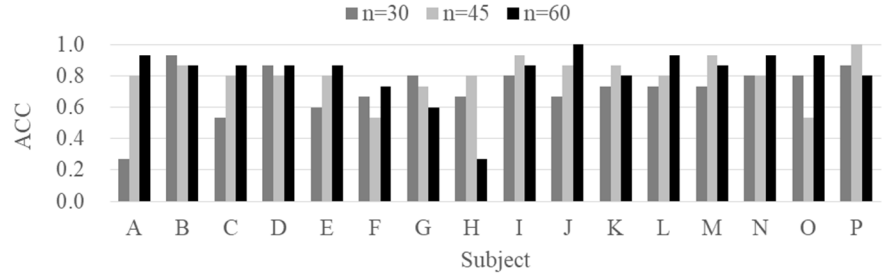
Next, to evaluate generalization capability with respect to inter-individual variation for recognizing objects, leave-one-out cross-validation (LOOCV) was applied on the data of 16 subjects. LOOCV uses one subject as validation data, and use the remained data as training data. The obtained recognition rates are shown in Fig. 3.11. It shows that the correct recognition was obtained for all combinations with printed images (O2) and still monitor images (O3) at the evaluation time length n = 60 frames; with other settings of evaluation time, recognition rates were 80% or higher.



(a) Printed image(O2).



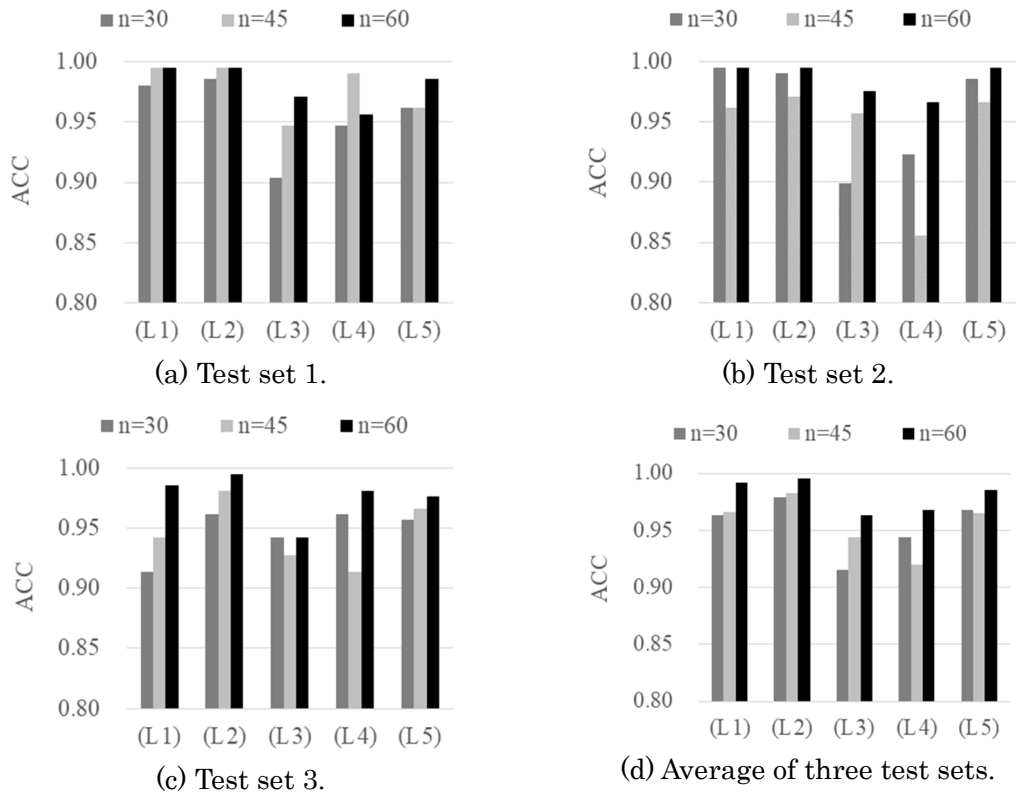
(b) Monitor image(O3).



(c) Monitor video(O4).

Fig. 3.11 Comparison of accuracy among subjects (A-P).

To examine the effect of illumination conditions, real subject data collected at 5 environments were used. The spoofing objects were 3D objects (O1: dolls) and plane objects (O2: printed images and O3: still monitor images). Videos (O3) were excluded because both R-G correlation and inter-area correlation feature values showed the same trends as real subjects. Data were repeatedly collected 3 times for every condition, and three-fold cross-validation was applied. The obtained results are presented in Fig. 3.12. As indicated by Fig. 3.12(d), accuracies under room light (L1) and front sunlight (L2) were 96.3% and 97.9% respectively at n = 30 frames, and 99.2% and 99.5% respectively at n = 60 frames. Under back sunlight (L3-L5), accuracies were 91.5%, 94.4%, 96.8% at n = 30 frames, and 96.3%, 96.8%, 98.6% at n = 60 frames.



(a) Test set 1.

(b) Test set 2.

(c) Test set 3.

(d) Average of three test sets.

Fig. 3.12 Comparison of accuracy among lighting conditions (L1-L5).

3.5 Discussion

3.5.1 Limitations of extracted features

For spoofing with videos, because both R-G correlation and inter-area correlation feature values showed the same trends as real subjects, spoofing detection proved difficult. The reason might be the similar blood perfusion changes for videos played on monitor and data of real subjects as shown in Fig. 3.7(d).

There is a possibility to increase the inter-area correlation feature by subject's movement. To address this issue, we examined the relationship between frame-to-frame shift in template matching of the subregion B and both correlation features. Fig. 3.13 shows the correlation of both features with an average shift (during evaluation time n= 60 frames) for all experimental data. It shows that inter-area correlation for printed images and still monitor images increases with shift and approaches the positive correlation it is the trend of real subjects. However, R-G correlation feature showed no correlation. Therefore, it can be said that, although inter-area correlation increases with subject's movement, liveness detection can be performed using R-G correlation feature.

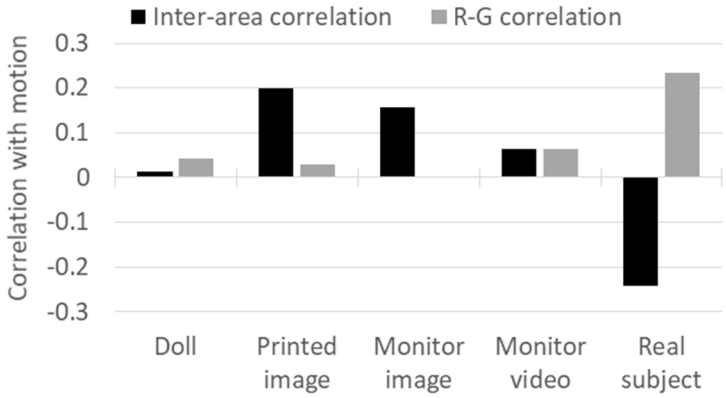


Fig. 3.136. Correlation with motion

3.5.2 Liveness detection results

In Fig. 3.10, the accuracy rate was 90% or higher for all evaluation data, except for videos (O4), at any evaluation time length. Particularly, with n = 60 frames, average accuracies for dolls, printed images, and still monitor images were 99.6%, 100%, and 98.3%, respectively. It shows that R-G correlation and inter-area correlation features are effective for detecting spoofing objects. In the evaluation for each subject at n = 60 frames, all of spoofing are correctly detected for all combinations (Fig. 3.11). Therefore, the proposed method is robust against interindividual variation.

As shown in Fig. 3.12(d), accuracy at the evaluation time length n = 30 frames was 96.3% for room light (L1) and 97.9% for sunlight in front of face (L2); at n = 60 frames, respective values were 99.2% and 99.5%. For the sunlight at the back of the head (L3-L5), the accuracy decreased. However, with the increase in evaluation time, the accuracy increased.

The accuracy reached as high as 96.3%, 96.8%, and 98.6% when the evaluation time was 60 frames. The evaluation time length, n, is the primary parameter determining response time. It should be set as small as necessary to obtain the required accuracy. Another factor related to response time is the window width ($2w$ + 1) of first-derivative Gaussian function. This was set as small as 11 frames (0.37 sec). Hence the response time is not affected by the ROI extraction process as ROI needs to be extracted from every frame.

3.6 Summary

Several methods were already proposed and verified to detect spoofing with photographs and other plane spoofing objects. However, there were hardly any simple and sufficient techniques to deal with 3D spoofing. In this chapter, we considered a fast recognition method of live subjects based on blood perfusion changes in time-series facial images. With 2-sec observation after frontal face identification, recognition accuracies of three types of spoofing objects (dolls, printed images, still monitor images) and real (human) subjects were 99.2% under room light (L1), 99.5% in case of sunlight in the front of the face (L2), and 98.6% for sunlight back of the head (L5). In evaluation using dolls, printed images, and still monitor images, recognition rates were, 99.6%, 100%, and 98.3% respectively. Therefore, the proposed method validated effective for 3D spoofing attacks as well. Since the proposed method uses time-series changes of green and red intensity values, a light source is necessary that has both green and red wavelengths of the light.

Limitation of the proposed method is that it could not detect spoofing with videos as the blood flow response was similar to the blood flow on human face. In the future, we will explore new features of video images that decline the image quality caused by re-recording. Also, the camera used in these experiments had relatively high sensitivity and could capture small details. Therefore, the effects of CMOS sensor sensitivity must be considered for practical application. Furthermore, subjects with diverse facial characteristics, such as bearded persons, will be considered.

Chapter 4 : Detecting pelvic fracture on 3D-CT using deep convolutional neural networks with multi-orientated slab images

4.1 Introduction

Pelvic fracture can be considered as a significant health concern, representing one of the most common causes of hospitalization and mobility loss [33]. Moreover, pelvic fracture is a key cause of mortality in the elderly [33][34][35][36]. The number of patients with pelvic fracture is continuously increasing among elderly populations in various countries, including Japan and the United States [37][38][39][40][41]. Quick and precise diagnosis is required in the hospital, especially in emergency departments, to enable early surgical intervention and preservation of the functionality of joints and quality of life [42][43]. This increase in patients is leading to an increasing load on radiologists, contributing to initial misdiagnoses [22]. Such misdiagnoses result in worsened prognosis, increased costs of treatment, and elevated mortality rates [34][45].

Pelvic fractures are more perceptible on images from computed tomography (CT) [46][47], which are widely used to diagnose pelvic fractures. As CT data usually contain a large number of images, a substantial investment in time is required to interpret each of the images to identify fractures, which then carries a risk of overlooking fractures [48]. An automated system to detect pelvic fractures from CT may thus assist physicians to diagnose fractures. Further, such results can be applied to augmented reality (AR) to assist surgeons in complex surgical procedures [49].

Several methods have been proposed to automatically detect pelvic fractures on CT. Chowdhury *et al.* [50] introduced some methods of pelvic fracture detection based on graph cut theory, curvatures, morphological analysis, and their combinations. It detected fractures by evaluating discontinuities or gaps in the pelvic bone. However, natural gaps exist between pelvic bones and could be incorrectly detected as fractures and thus increase the number of false-positive results. Another method was proposed to detect fractures on CT images of traumatic pelvic injuries based on the registered active shape model and 2D stationary wavelet transform [51].

Accuracy, sensitivity, and specificity in 12 subjects were 91.98%, 93.33%, and 89.26%, respectively. That method focused only on completely displaced bone fractures, and did not discuss incompletely displaced fractures or compression fractures. The number of subjects was also limited.

Some studies have detected various kinds of bone fractures on 2D X-ray radiographs based on deep convolutional neural networks (DCNNs). Lindsey *et al.* [52] proposed a method of wrist fracture detection. It estimated a conditional probability map which represents a probability of fracture at each pixel. Thian *et al.* [53] proposed a method to detect wrist fractures using frontal or lateral X-ray radiographs based on faster region-based convolutional neural network (Faster R-CNN) architecture. Detection accuracies in frontal and lateral radiographs of the wrist were 88.9% and 91.2%, respectively. A method to detect intertrochanteric hip fractures from X-ray radiographs of the femoral head and the greater and lesser trochanters was proposed based on VGG16 [54], a kind of DCNN. Detection accuracy was reported as 95.5%, higher than the detection accuracy of orthopedic surgeons (92.2%). Sato *et al.* [55] introduced a CNN based method to detect hip fracture on plain X-ray radiograph. The experimental results from 300 images showed that the accuracy, sensitivity, specificity, F-value, and area under the curve (AUC) were 96.1%, 95.2%, 96.9%, 0.961, and 0.99, respectively. Cheng *et al.* [56] developed a human-algorithm integration system to improve the diagnosis of hip fracture. Another method to classify proximal femur fracture from X-ray images was proposed based on a multistage architecture of successive CNNs in cascade along with gradient class activation maps (Grad-CAM) to visualize the most relevant areas of the images [57]. Mean accuracies of the method for 3-class and 5-class classifications were 0.86 and 0.81 respectively. The proposed CAD system based on the method improved accuracy of specialists by 14%. However, these methods were based on 2D images, and could not be applied directly to 3D images.

As related studies involving fractures at sites other than the pelvis, a few studies have proposed methods of automated bone fracture detection using CT. Bar *et al.* [58] proposed a method to detect vertebral compression fractures (VCFs), based on DCNN and long short-term memory (LSTM). This method first estimated a vector of probabilities from patches of CT images using DCNN, then classified these patches into VCF using LSTM. Accuracy, sensitivity, and specificity were 89.1%, 83.9%, and 93.8%, respectively. Roth et al. [59] proposed a method to detect posterior element fractures from CT images based on ConvNet. Sensitivities at 5 false positives per

patient (FP/P) and 10 FP/P were 71% and 81%, respectively. Recently, Zhou *et. al.* proposed an automatic method to detect and classify rib fractures on thoracic CT [60]. This method is based on Faster R-CNN. With the assistance of this method, the sensitivity for diagnosing rib fractures was increased by 23.9%.

Those papers mainly focused on detecting bone fractures in 2D spaces. The methods used CT, but did not segment 3D fracture regions, and did not consider 3D image features and structure. Basically, those methods cannot evaluate the 3D spatial connectivity of fractures. A straightforward approach to evaluating 3D information is 3D-DCNN [61], but as the availability of 3D data is limited, a suitable 3D-DCNN model to detect fractures is not yet known, and the method would be computationally costly. Another approach is to synthesize 2D images from 3D volume data, known as 2.5D representation. Such 2.5D representation has been applied to lymph node detection using CT images [62] and cerebral aneurysm detection on Magnetic Resonance (MR) angiography [63]. The 2.5D approach synthesizes 2D images from 3D volume data in orthogonal and diagonal directions. The synthesized 2D images may contain a large amount of 3D information in comparison with the original raw 2D images.

Bone fractures on CT images can take various appearances. Any surface displaced due to bone fracture and vertical to the imaging plane will be clearly apparent. However, fracture surfaces displaced parallel to the imaging plane can be hard to recognize. This means that appropriate orientation of the imaging plane is crucial. However, the appropriate orientation cannot be determined initially, because each fracture has a different orientation, and acquisition of images in multiple orientations from the same patient is unfeasible because of the risks associated with X-ray exposure. The present study addressed this obstacle by reconstructing raw sectional images into multiple-orientated images. We assume that detection accuracy in 3D space would thus be improved by detecting fractures in the reconstructed multiple-orientated images simultaneously and aggregating those in 3D space.

This chapter proposes a fully automated method of fracture detection on 3D- CT of pelvic region. The proposed method is based on multiple 2D-DCNNs, in which each 2D-DCNN evaluates images in a different orientation. This utilizes YOLOv3 [64], a real-time object detection system, to detect fractures on 2D images. For each orientation, three 2.5D slab images are synthesized with three different thicknesses. Fracture candidates are detected by each YOLOv3 model with different orientations

simultaneously. The 3D fracture region is finally detected by integrating fracture candidates. By detecting bone fractures in multiple orientations, the proposed method improves detection accuracy.

4.2 Subjects and materials

CT images taken at the Steel Memorial Hirohata Hospital were used in this study. Approval was obtained from the Institutional Ethics Committee (IRB #1-52, Steel Memorial Hirohata Hospital), and the need to obtain informed consent from subjects was waived. These analyses were performed in accordance with the relevant rules, guidelines and regulations.

Two datasets were used in this study. Dataset **A** consists of CT images acquired from 93 subjects who had one or more pelvic fractures. Dataset **B** consists of CT images acquired from 112 subjects identified by orthopedic surgeons as not having any fractures. Both datasets were acquired at Steel Memorial Hirohata Hospital, Japan.

Dataset **A** was taken from 47 male and 46 female subjects with a mean age of 66.1 ± 18.9 years (range, 20–93 years). Each subject had one or more fractures of the pelvis, and no implant had been confirmed on CT images. Before subjects received surgical treatment, CT images were acquired using three multidetector-row CT (MDCT) scanners (SOMATOM Definition AS 32 line, SOMATOM Go. Top 64 line, or Sensation Cardiac 16 line; Siemens, Germany). The images were taken between April 2013 and August 2019. CT images covered the whole pelvis, and image acquisition parameters were: tube voltage, 120 kVp; current, auto mAs; spatial resolution, 0.61–0.98 mm; and thickness, 0.60–1.00 mm. No space between slices was used. All CT images were annotated by orthopedic surgeons for training and evaluation purposes. The annotation procedure is described in the following section. Dataset **A** was used for both training and evaluation.

Dataset **B** was taken from 69 male and 43 female subjects with a mean age of 61.3 ± 19.7 years (range, 20–93 years). No fractures or implants were confirmed on CT images by orthopedic surgeons specializing in pelvic fracture. CT images were acquired between July 2018 and December 2018 using an MDCT scanner (SOMATOM Definition AS 32 line; Siemens, Germany). CT images covered the whole pelvis, and image acquisition parameters were: tube voltage, 120 kVp; current, auto mAs; spatial resolution, 0.61–0.98 mm; and thickness, 0.70 mm. No space between

slices was used. Dataset **B** was used for evaluation only.

The acquired CT images had 12-bit pixel resolution. As a preliminary step, CT values of 1–1,800 HU were linearly converted into 0–255. To normalize pelvic size, CT images were normalized into 296×169×288 mm, as the average size of the pelvis for 30 randomly selected subjects, using B-spline interpolation. The resulting dimensions were 494×282×480 voxels, and voxel size was 0.6×0.6×0.6 mm.

4.3 Proposed Method

4.3.1 Overviews

A conceptual diagram of the proposed method is illustrated in Fig. 4.1. The method first synthesizes 2.5D slab images with thicknesses of 18.6 mm, 9.0 mm, and 0.6 mm in nine orientations (Fig. 4.1b) from the provided CT images (Fig. 4.1a). Second, the method detects fracture candidates for each orientation using YOLOv3 model simultaneously (Fig. 4.1c). Third, 3D volumes of fracture candidates are formed by thickening the detected 2D boundary box (Fig. 1d). Finally, the 3D fracture region is determined by integrating fracture candidates (Fig. 4.1e).

The method has two parameters, C_{th} and I_{th}. C_{th} represents a threshold of confidence score to detect the 2D bounding box with YOLOv3 model. The confidence score takes a value between 0 and 1, with higher values showing higher confidence. Bounding boxes with confidence scores equal to or exceeding C_{th} are detected. I_{th} represents the threshold for the degree of fracture. When the number of orientations in which the voxel is included in fracture candidates equals or exceeds this threshold, the voxel is extracted as a fracture voxel.

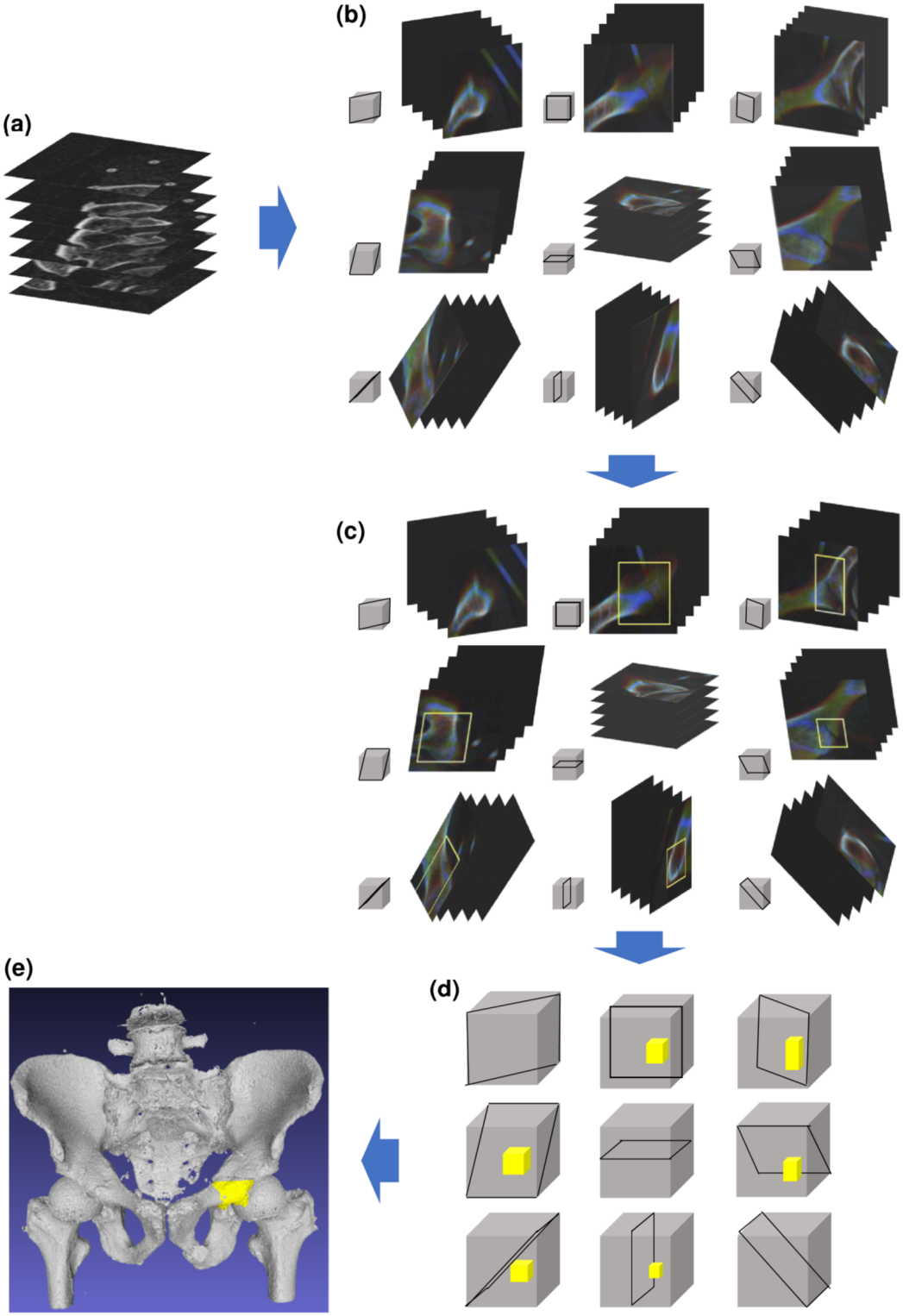


Fig. 4.1 Conceptual diagram of the proposed method. (a) A series of axial CT images obtained from a subject. Each image represents 50×50-mm area for easy understanding. (b) Nine synthesized, orientated 2.5D images. Three slab images with thicknesses of 18.6 mm, 9.0 mm, and 0.6 mm are visualized by R-G-B colors, respectively. (c) Detection of 2D fracture candidates. (d) Thickening of 2D fracture candidates. (e) Fracture region detection.

4.3.1 Multi-orientated image synthesis

A volume of a subject is first divided into four cubes to reduce the memory required for YOLOv3 analysis. The dimensions of each cube are 282×282×282 voxels. Cubes are extracted from the edge of the normalized volume so that overlaps between cubes are minimal. The proposed method synthesizes 9 orientation images from each cube; the 9 orientations are 3 orthogonal directions and 6 diagonal directions. In the synthesized images, the out-of-imaging area is filled by 0. Next, for each sectional image, three slab images with 31-image (18.6 mm), 15-image (9.0 mm), and 1-image (0.6 mm) thickness are synthesized to represent neighboring information according to 2.5-D representation. Slab images are synthesized by averaging neighboring images. Fig. 4.2 shows an example of three slab images.

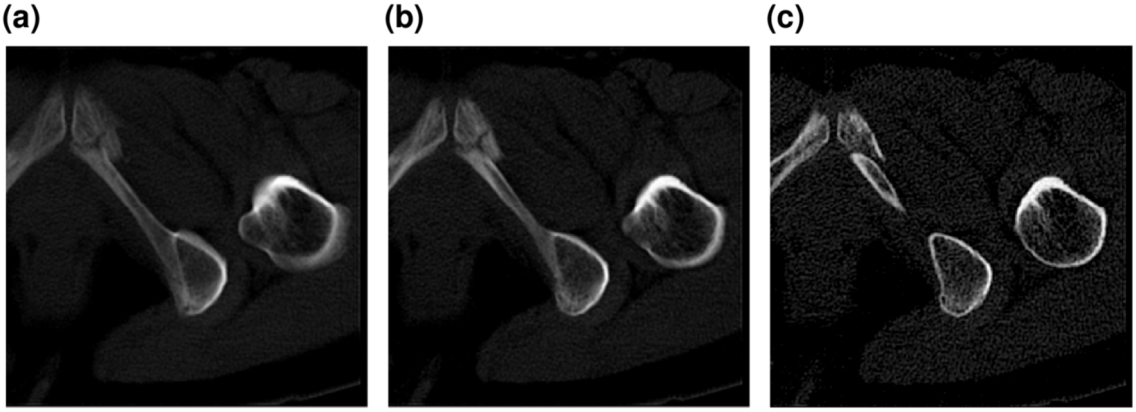


Fig. 4.2 A 2.5D representation. (a) Image with 31-image thickness. (b) Image with 15-image thickness. (c) Image with 1-image thickness.

4.3.2 Bone fracture region extraction method

The proposed method simultaneously extracts 2D bone fracture candidates from multiple 2D orientation images. By aggregating 2D bone fracture candidates, the 3D bone fracture region is segmented. The procedure for bone region extraction is described below.

Step 1 Detect bone fractures from 2D images using multiple YOLOv3 models in parallel. Nine YOLOv3 models are prepared to analyze 9 orientation images. For each orientation, a set of three thicknesses of slab images is fed to the input layer of YOLOv3 model, which then yields coordinates of bounding boxes and confidence scores. If the confidence score is greater than or equal to a threshold confidence score

(C_{th}), the bounding box is detected as a fracture candidate. The fracture candidates are thickened by 12.6 mm to cover the whole fracture volume, such as that for a completely displaced fracture.

Step 2 Integrate the thickened fracture candidates detected from the multiple YOLOv3 models into one 3D volume. Each voxel represents the degree of fracture, defined as the number of orientations included in the thickened fracture candidates. Each voxel takes a value between 0 and 9.

Step 3 Segment the fracture region by thresholding the obtained 3D volume. The voxels with a value equal to or higher than I_{th} are set to 1, and all others are set to 0. Small fracture regions in which the number of voxels is less than 8,000 are discarded to suppress over-detection. The remaining regions are finally detected as fracture regions.

4.3.4 New 3D surface annotation method

The proposed method requires annotation of bone fracture regions to train YOLOv3 models. However, the number of CT images is huge, and manually performing the annotation procedure that surrounds a fracture area with a polygon is too difficult. To annotate fractures efficiently, this study introduces a new 3D annotation scheme using 3D surface rendering. The 3D surface rendering is performed by representing the pelvic bone surface on CT images with a set of small polygons. The pelvic bone region is easily segmented using image processing such as thresholding, morphological operation, etc. Orthopedic surgeons select 3-4 adjacent polygons around fractures using the 3D surface rendering as shown in Fig. 4.3a. For example, a completely displaced fracture (**F1**) is annotated as shown in Fig. 4.3b. An incompletely displaced fracture (**F2**) or a compression fracture (**F3**) is annotated as show in Fig. 4.3c. After 3D annotation on the pelvic bone surface, the annotated 3D polygons are converted into 2D bounding boxes on sectional images for each orientation. Because YOLOv3 model evaluates the three slab images with 18.6 mm, 9.0 mm, and 0.6 mm thicknesses, the 2D bounding boxes are also thickened by 18.6 mm.

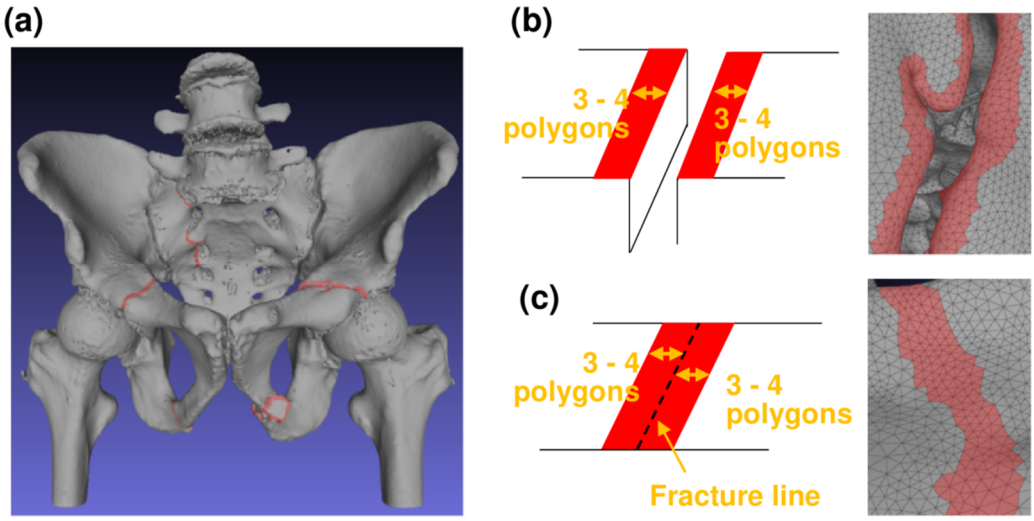


Fig. 4.3 The 3D annotation method. (a) Annotated 3D bone surface data. (b) Annotation of completely displaced fracture (**F1**). (c) Annotation of incompletely displaced fracture (**F2**) or compression fracture (**F3**).

4.3.5 YOLOv3 model training

A YOLOv3 model [64] pre-trained with the 'COCO trainval' dataset is used. YOLOv3 model is trained using a set of three 2.5D images and the ground truth bounding boxes. A different YOLOv3 model is trained for each of the 9 orientation images, and 9 YOLOv3 models are obtained. The training data are augmented by brightness adjustment, rotation, horizontal flip, enlargement, reduction, and changing the aspect ratio. Each model is fine-tuned on three output layers for the first eight epochs with a learning rate of 0.001, then all layers are fine-tuned for the following ten epochs with a learning rate of 0.0001. The batch size is 28. The input size of the model is 416×416.

To correct the imbalance in the number of images with and without fractures, the volume of a subject is divided into volumes with 20 consecutives slices. For each divided volume, when the number of images with fractures over all subjects is less than 10%, 10% of images without fracture from the same volume are randomly chosen. Otherwise, the same number of images without fractures are selected randomly. For training the model, the multi-orientated synthesized images from dataset **A** are used. Table 4.1 shows the total number of synthesized images from dataset **A** for each orientation. The data were decomposed into 6 folds to perform 6-folds-cross-validation test.

Table 4.1 Total number of synthesized images from dataset **A** (93 subjects).

Orientations	with fracture	without fracture
orthogonal direction 1	31,217	62,527
orthogonal direction 2	23,524	70,220
orthogonal direction 3	29,766	63,978
diagonal direction 1	29,298	107,970
diagonal direction 2	33,942	103,326
diagonal direction 3	28,011	109,257
diagonal direction 4	30,385	106,883
diagonal direction 5	32,697	104,571
diagonal direction 6	29,083	108,185

4.4 Experimental Results

4.4.1 Evaluation metrics

Precision, recall, F-score, and AUC are calculated to evaluate the results. The ground truth of the 3D fracture region is prepared by intersections of 2D fracture boundary boxes at every orientation image. Then, IoU is calculated between the detected and ground truth 3D fracture regions. The IoU is defined by equation (4.1).

$$IoU = \frac{A_m \cap B_n}{A_m \cup B_n}, \tag{4.1}$$

where A_m is a set of ground truth fracture regions, and B_n is a set of the detected fracture regions. The correspondence between ground truth and detected region is determined by maximizing IoU. When IoU is greater than or equal to a threshold, the ground truth region is successfully detected. Otherwise, the ground truth region is not detected.

Fracture-wise precision and recall are calculated using true positive (TP), false positive (FP), and false negative (FN). TP denotes the number of ground truth fractures successfully detected. FP denotes the number of fractures detected incorrectly. FN denotes the number of ground truth fractures that are not detected. Precision, recall, and F score are defined by equations (4.2)-(4.4).

$$Precision = \frac{TP}{TP+FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.3}$$

$$F\ score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{4.4}$$

The interpolated precision [65] is calculated by sampling precision whenever it drops and computing the sum of the rectangular blocks using equation (4.5).

$$p_{interp}(r_n) = \max_{\tilde{r}:\, \tilde{r} \geq r_n} \left(p(\tilde{r})\right), \tag{4.5}$$

where $p(r)$ = precision at recall r.

4.4.2 Detection of 3D fracture regions

Fig. 4.4a shows the estimated degree of fracture overlaid on multiplanar reconstruction images with C_{th} of 0.2. The degree of fracture is estimated at each voxel, and assumes a value between 0 and 9 as the number of orientations under evaluation; 0 means that no fracture is detected in any orientation, and 9 means that fracture is detected in all orientations. Fig. 4.4b shows the resultant fracture region with I_{th} = 6. Over-detection occurring in the individual orientation detection step is suppressed by aggregation of fracture candidates for each orientation.

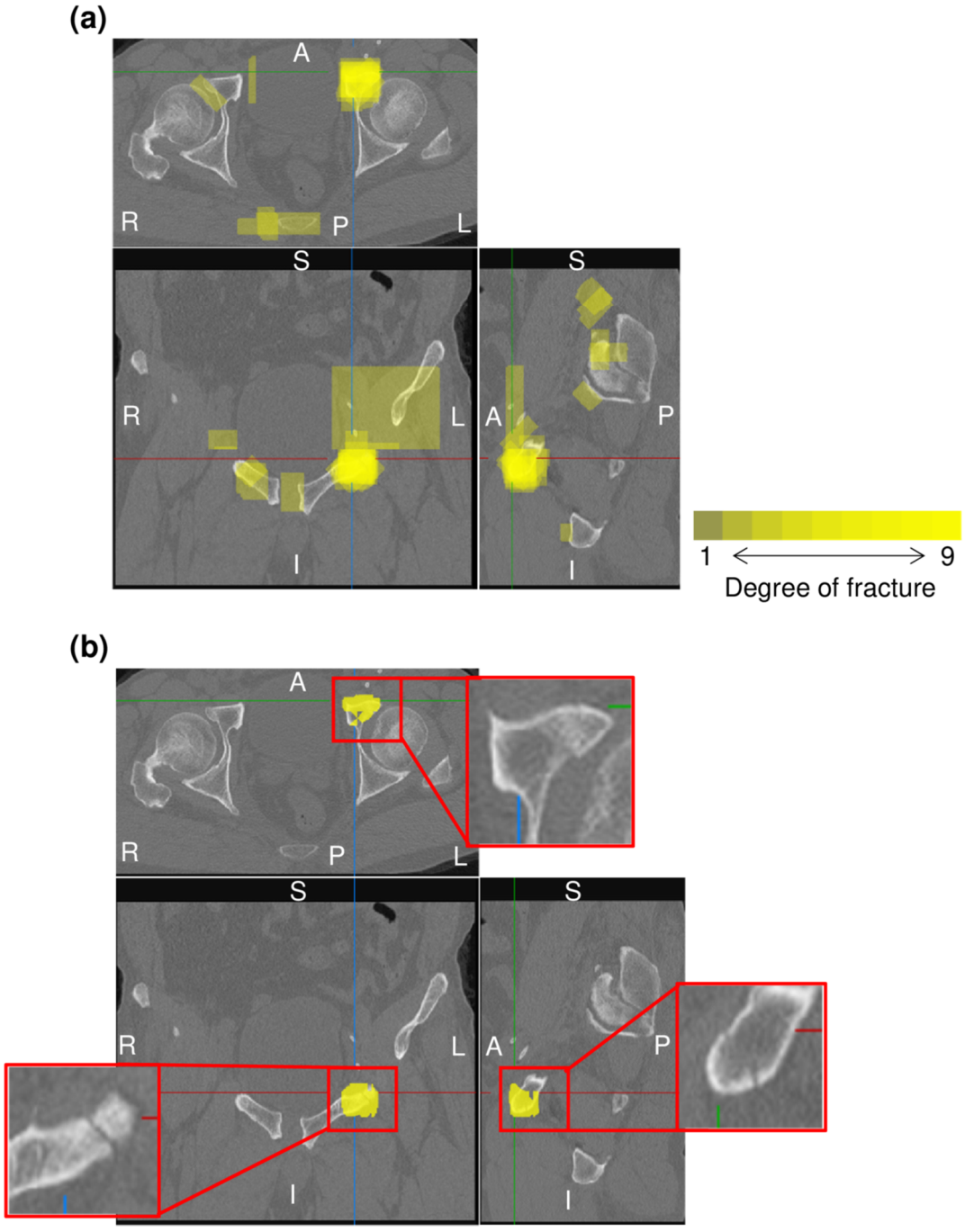


Fig. 4.4 Estimated degree of fracture on multiplanar reconstruction images. Top: axial image; bottom-left: coronal image; bottom-right: sagittal image. L: left; R: right; A: anterior; P: posterior; S: superior; I: inferior. (a) Integrated 3D fracture candidate region overlapping on CT images (C_{th}: 0.2). Yellow represents the degree of fracture. (b) Resultant 3D fracture region (I_{th}: 6). Yellow represents the detected region. The enlarged image shows raw CT images for the detected fracture region.

4.4.3 Detection accuracy

Dataset **A** with bone fractures was used to evaluate the proposed method, and 6-fold cross-validation was conducted. Dataset **A** included 93 subjects with 389 fractures. All subjects were divided into 6 groups, with 5 groups used for training, and the remaining group used for evaluation. To evaluate performance, the IoU between the detected fracture region and ground truth fracture region was calculated. Detection accuracy was evaluated for each fracture, and the evaluation metrics were precision, recall, and AUC. The threshold used for IoU was 10%. Fig. 4.5 shows the interpolated precision-recall (PR) curve, obtained using the set of parameters as combinations of C_{th} = 0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8 and I_{th} = 3, 4, 5, 6, 7, 8, 9. The AUC for multiple orientations was 0.824 with an IoU of 10%.

To demonstrate the effectiveness of the proposed method using multiple orientations, single-orientation detection, and triple-orientation results are also plotted in Fig. 4.5. The single-orientation method detected fractures using only axial images, and the triple-orientation method used axial, coronal, and sagittal images. AUCs with the single- and triple-orientation methods were 0.652 and 0.734, respectively. The proposed method detected bone fractures successfully using more orientations, and the less-orientated method failed when the fracture did not appear clearly in the given orientation. We concluded that multiple-orientated analysis is quite effective to detect bone fractures from CT images.

Parameters C_{th} and I_{th} should be optimized to provide the highest value of the two evaluation metrics “recall” and “precision”, although a tradeoff exists between recall and precision. F score was therefore used to evaluate the overall performance of the proposed method. The highest F score for an IoU of 10% was 0.853 when C_{th} was 0.2 and I_{th} was 6. Recall was 0.805 and precision was 0.907.

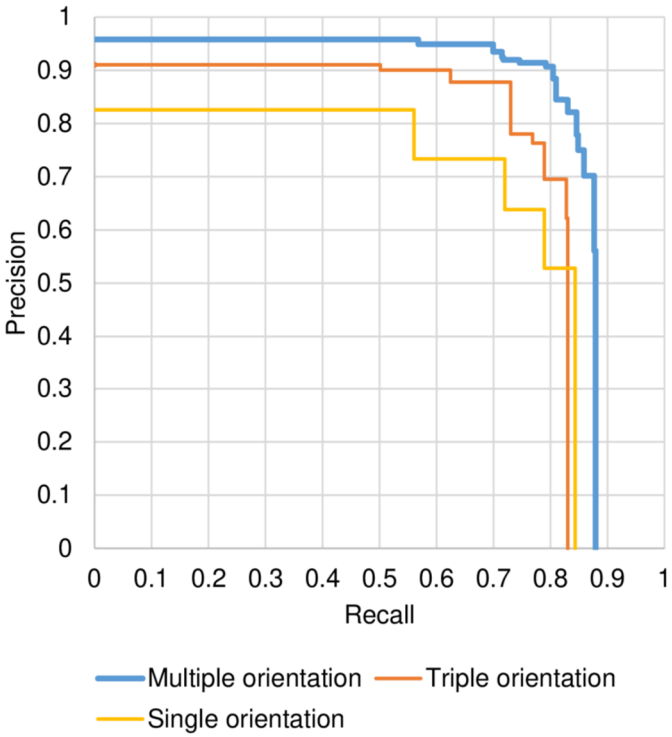


Fig. 4.5 Precision-recall curve.

4.4.4 3D visualization of the detected fractures

Fig. 4.6 shows the comparison of ground truth fractures and automatically detected fractures under the proposed method. Fractures on the 3D bone surface are highlighted. The subject had five ground truth fractures (**A**-**E**). The proposed method successfully detected all except one fracture (**C**). The IoUs of **A**-**E** were 15.4, 7.3, 0.0, 30.3, and 15.0, respectively. Although the detected volume is slightly different from the ground truth fractures, the detected region is located close to the ground truth fractures, and will assist physicians in identifying the fractures.

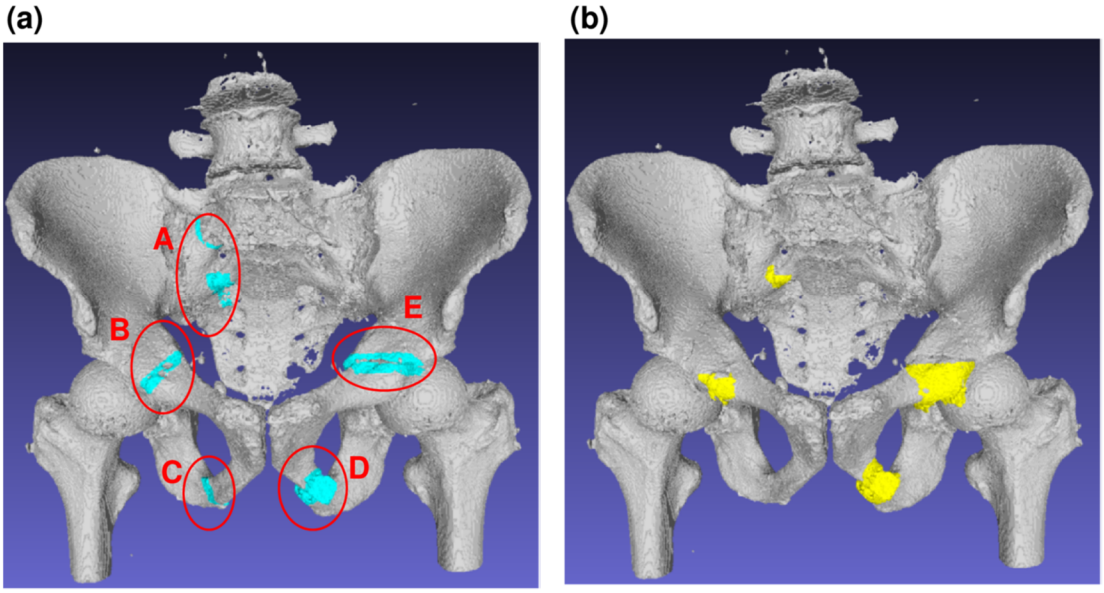


Fig. 4.6 The 3D visualization of fractures. (a) Ground truth fractures. (b) Automatically detected fractures.

4.4.5 Subject-wise recall and specificity

Subject-wise recall and specificity were evaluated using dataset **A** with bone fractures, and dataset **B** without bone fractures, respectively. For comparison, we called the recall and specificity calculated for each fracture the fracture-wise recall and specificity. Only dataset **A** was used for training YOLOv3 models, and 6-fold cross-validation was also conducted. Analysis parameters were: C_{th} was 0.2, I_{th} was 6, and the threshold of IoU was 10%. Subject-wise recall and specificity were evaluated for each subject (not for each fracture), where a positive subject denotes a subject in whom one or more fractures are detected, and a negative subject denotes a subject in whom no fractures are detected. Subject-wise recall calculated using dataset **A** was 1.00 (93 of 93 subjects), showing that the proposed method completely detected all subjects with bone fractures. The ratio of subjects for whom all fractures were detected was 0.559 (52 of 93 subjects). Subject-wise specificity for dataset **B** was 0.964 (4 of 112 subjects), and the proposed method successfully recognized all except 4 non-fracture subjects.

4.5 Discussion

The experimental results of the proposed method depend on parameters, C_{th} and I_{th}. Fig. 4.7a shows F score at IoU of 10% with changes in C_{th} and I_{th}. For each of

C_{th}= 0.4, 0.2, 0.1, and 0.02, the highest F score was 0.849 (recall 0.792, precision 0.914) at I_{th} = 5, 0.853 (recall 0.805, precision 0.907) at I_{th}= 6, 0.845 (recall 0.810, precision 0.884) at I_{th}= 7, and 0.832 (recall 0.802, precision 0.863) at I_{th}= 8. A tendency was seen for I_{th} to be decreased when C_{th} was large, while the I_{th} should be increased when C_{th} is small. This is because that the number of fracture candidates detected by multiple YOLOv3 models in parallel increases with decreasing C_{th}, and can be suppressed by increasing I_{th} at the integration step. Next, Fig. 4.7b shows a cumulative histogram of IoUs of the detected fractures by the proposed method with I_{th}= 5, 6, and 7. This shows that the ratio of high IoU fractures increased with higher I_{th}, because the integration of multiple orientation results specifies the fracture region more precisely.

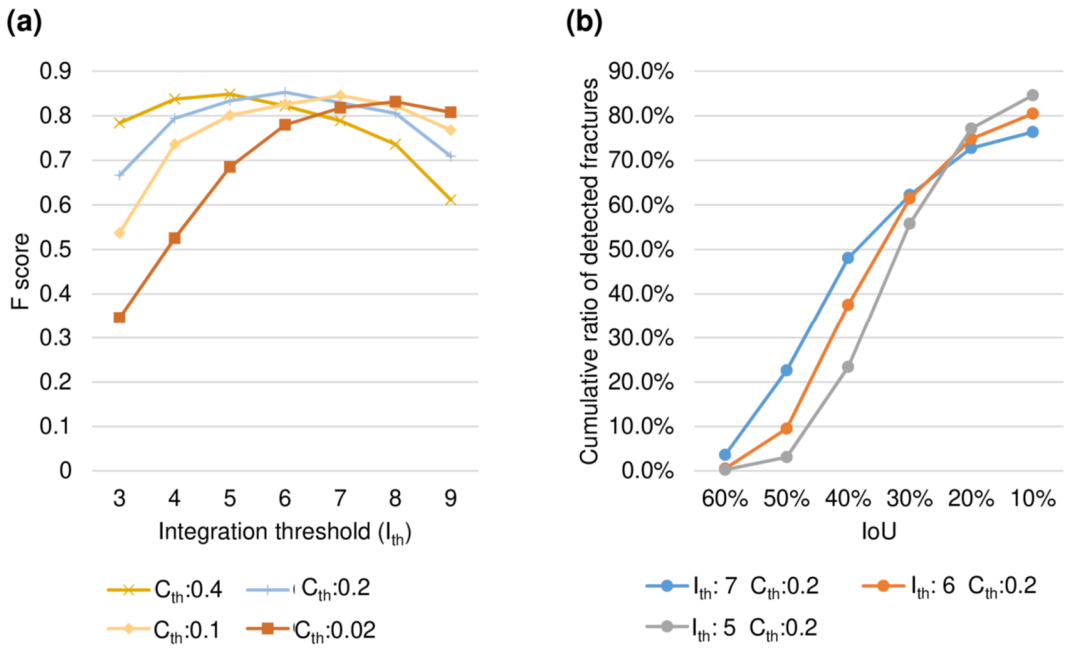


Fig. 4.7 Performance dependency on analysis parameters. (a) Relationship between I_{th} and C_{th}. (b) Cumulative histogram of IoU of the detected fractures.

Next, detection accuracy among appearance types was discussed. We classified bone fractures into 3 types: (**F1**) completely displaced fracture; (**F2**) incompletely displaced fracture; and (**F3**) compression fracture. **F1** type represents fractures where fractured part of the bone is completely separated (Fig. 4.8a). **F2** type represents fractures where the fractured part of the bone is loosely separated (Fig. 4.8b). **F3** type represents the fractures where the fractured part of the bone is not

separated but a part of the bone surface has changed (Fig. 4.8c). The 389 fractures of dataset **A** were classified into 67 **F1** fractures, 282 **F2** fractures, and 40 **F3** fractures. Fig. 4.8 shows examples of axial CT images for the 3 fracture types. We calculated fracture-wise recall for each type of fracture using parameters with C_{th} of 0.2 and I_{th} of 6 that provided the highest F score. The fracture-wise recalls were 0.955 (**F1**), 0.869 (**F2**), and 0.350 (**F3**). The accuracy of **F3** type was lower than that of the other types, because few characteristics of fracture were present on the image.

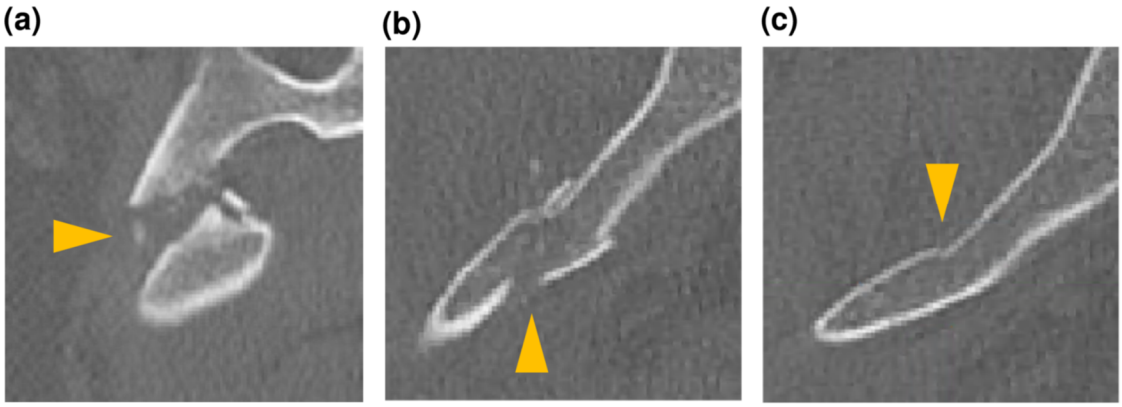


Fig. 4.8 Types of fractures. (a) Completely displaced fracture (**F1**). (b) Incompletely displaced fracture (**F2**). (c) Compression fracture (**F3**). Fractures are indicated by triangles.

Processing time for one subject was 756 sec in total, using a computer with an i9-10900k CPU, and a TITAN-RTX GPU. The method consists of 3 steps; the first step to synthesize slab image took 361 sec, the second stage to predict 2D fracture candidates took 253 sec, and the third stage to aggregate them into 3D space took 142 sec.

4.6 Summary

Although many studies have been conducted using DCNN to detect lesions in 2D and 3D medical images, training DCNN using 3D images is complicated, computationally expensive, and requires a large amount of training data. Therefore, we have trained multiple YOLOv3 models in parallel. In the proposed method, each YOLOv3 model was trained using multiple oriented 2D slab images constructed from 3D-CT. We assumed that there is an appropriate orientation to extract maximum features for each fracture. For this reason, multiple YOLOv3 models were used to detect 2D fracture candidates in different orientations. The 3D fracture region was

detected by integrating the 2D fracture candidates. The proposed method was tested on 93 subjects with fractures. The area under the curve (AUC), recall, and precision were 0.824, 0.805, and 0.907 respectively. To evaluate the specificity, we tested our method on 112 subjects without fractures. The specificity was 96.4% and fractures were detected on only 4 of the 112 subjects.

The proposed method will assist physicians to detect pelvic fractures. While fracture detection performance will be increased, the risk of misleading physicians must be considered. Use of the method should thus be limited to second-stage interpretations after the first interpretation without the AI-system. A limitation to the proposed method is that it is not applicable to patients with implants. The future prospects for the proposed method include extending the methods for patients with implants, compared with other object detection methods such as Faster R-CNN, SSD, and optimization of deep learning parameters.

Chapter 5 : Conclusion

In this dissertation, high-dimensional biomedical image recognition methods have been proposed that incorporate a new analysis perspective with the goal of saving labor in human operations on high-dimensional biomedical images that have been taken in large numbers in the fields of medical care and security in recent years.

In Chapter 2, a multiple-measurement-points-voting-method has been proposed to estimate the RRI from the most frequent value of the RRI calculated for each pixel in the face area, which is the variation of the period per heart beat (RRI: R-R interval) that is related to the state of the autonomous nervous system. The method was evaluated using data in a stationary state (198 data, 3 minutes per data) and data in a working state (60 data: 2 minutes per data), and showed that the proposed method was significantly more accurate than the conventional method of ICA for both data sets. In addition, the proposed method reduced the observation time required for reliable RRI estimation by about 47% compared to the conventional method. Furthermore, a parallelization method was devised for the implementation of the proposed method, and achieved real-time processing for data with a ROI size within 302×384 pixel.

Chapter 3 has proposed a method to distinguish a real person in a short time by focusing on the changes in blood flow that appear on time-series facial images, which are unique to living bodies. After detecting a frontal face, recognition rates for indoor lighting, normal lighting, and lighting at the back with direct illumination on the face were 99.2%, 99.5%, and 98.6%, respectively using an observation time of 2 sec. In the evaluation of each type of spoofing objects, the recognition rates for dolls, printed photographs, and still images of monitors were 99.6%, 100%, and 98.3%, respectively, indicating that the proposed method was effective against spoofing attacks by 3D objects.

Chapter 4 proposes a novel method to detect factures in pelvic CT images. It uses 2D cross-sectional images of nine directions reconstructed from CT images to extract 2D fracture candidates using a deep convolutional neural network. Then, the detected 2D fracture candidates in nine directions were integrated in 3D space to form the 3D fracture candidate region. This procedure significantly reduced the computational cost because the images are analyzed by deep learning in 2D. In

addition, since the convolutional neural network is trained by using 2D cross-sectional images, we can prepare a large number of training images from a single subject. To annotate a large number of training images, this study introduced an automated method for annotating 2D cross-sectional images in each of the nine directions by using the fracture lines annotated on the 3D bone surface model. The proposed fracture detection method was validated on 93 subjects with fractures, and the AUC was 0.824, the recall was 0.805, and the precision was 0.907. It was also applied to 112 subjects without fractures, and 108 subjects (96.4%) were predicted correctly.

These studies discuss three kinds high-dimension of signals or images to improve the recognition performance in biomedical applications with artificial intelligence. At first, it showed that RRI, a biomarker for detecting physical condition, can be detected with higher accuracy and shorter observation time than conventional methods by measuring and voting at multiple points on the facial video. The first study discusses on high-dimension of RRI measurement signals. In addition, by comparing the intensity signal from the facial video image among three wavelengths and among multiple regions of interest, it was possible to detect spoofing by 3D objects, which is a problem in face recognition. The second study discusses on high-dimension of the blood flow measurement signal. Furthermore, 3D pelvic fractures with complex patterns can be detected by reconstructing the cross-section of CT images in multiple directions. The third study discusses on high-dimension of fracture detection cross-section. In this way, the range of artificial intelligence applications had been expanded by increasing the dimensionality of feature extraction from high-dimensional biomedical images.

Future work includes to proceed feasibility study of these researches, such as improving the accuracy of pelvic fracture detection, and extending the function to whole-body fracture detection.

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering”, Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 815-823, 2015.

[2] H. Fujita, “Now evolving and diversifying computer-aided diagnosis (CAD)”, Japanese J. of Imaging and Information Sciences in Medicine, Vol. 36, No. 2, pp. 25-29, 2019. (in Japanese)

[3] S. Okada and R. Ishii, “Social signal processing and AI”, Trans. of the Japanese Society for Artificial Intelligence, Vol. 32, No. 6, pp. 915-920, 2017. (in Japanese)

[4] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, “3D deep learning on medical images: a review”, Sensors, Vol. 20, No. 18, pp. 5097, 2020.

[5] K. Ukai, R. Rahman, and S. Kobashi, “Short-time estimation of R-R interval from facial video image with a multiple-measurement-points-voting-method”, Trans. of the Institute of Systems, Control and Information Engineers, Vol. 31, No. 12, pp. 403-411, 2018. (in Japanese)

[6] M. Pagani, F. Lombardi, S. Guzzetti, O. Rimoldi, R. Furlan, P. Pizzinelli, G. Sandrone, G. Malfatto, S. Dell’Orto, E. Piccaluga, M. Turiel, G. Baselli, S. Cerutti, and A. Malliani, “Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho–vagal interaction in man and conscious dog”, Circulation Research, Vol. 59, No.2, pp. 178-193, 1986.

[7] F. Toyofuku, K. Yamaguchi, and H. Hagiwara, “Simplified method for estimating parasympathetic nervous activity by Lorenz plot of ECG RR intervals”, The Japanese Journal of Ergonomics, Vol. 43, No. 4, pp. 185-192, 2007. (in Japanese)

[8] Y. Matsumoto, N. Mori, R. Mitajiri, and Z. Jiang, “Study of mental stress evaluation based on analysis of heart rate variability”, Journal of Life Support Engineering, Vol. 22, No. 3, pp. 19-25, 2010. (in Japanese)

[9] K. Takagi, “Plethysmography”, Transactions of the Japanese Society for Medical and Biological Engineering, Vol. 3, No. 1, pp. 3-14, 1965. (in Japanese)

[10] M. Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in noncontact, multiparameter physiological measurements using a wabcam”, IEEE Engineering in Medicine and Biology Society, Vol. 58, No. 1, pp. 7-11, 2011.

[11] H. Rahman, M. U. Ahmed, S. Begum, and P. Funk, “Real time heart rate monitoring from facial RGB color video using webcam”, Proceedings of 29th Annual Workshop of the Swedish Artificial Intelligence Society, pp.8-15, 2016.

[12] J. Ruminski, “The accuracy of pulse rate estimation from the sequence of face images”, Proceedings of 9th International Conference on Human System Interactions, pp. 518-524, 2016.

[13] K. Ukai, R. Rahman, and S. Kobashi, “Facial skin blood perfusion change based liveness detection using video images”, IEEJ Tran. on Sensors and Micromachines, Vol. 139, No. 2, pp. 29-37, 2019. (in Japanese)

[14] K. Ukai, R. Rahman, N. Yagi, K. Hayashi, A. Maruo, H. Muratsu, and S. Kobashi, “Detecting pelvic fracture on 3D-CT using deep convolutional neural networks with multi-orientated slab images”, Scientific Reports, Vol. 11, No. 11716, 2021.

[15] T. Ogasawara, K. Ono, N. Matsuura, M. Yamaguchi, J. Watanabe, and S. Tsukada, “Development of applications for a wearable electrode embedded in inner shirt”, NTT Technical Review, Vol. 26, No. 11, pp. 16-20, 2014. (in Japanese)

[16] T. Sakamoto, S. Okumura, R. Imanishi, H. Taki, T. Sato, M. Yoshioka, K. Inoue, T. Fukuda, and H. Sakai, “Remote heartbeat monitoring from human soles using 60-GHz ultra-wideband radar”, IEICE Electronic Express, Vol. 12, No. 21, pp. 1-6, 2015.

[17] C. Takano and Y. Ohta, “Heart rate measurement based on a time-lapse image”, Medical Engineering & Physics, Vol. 29, No. 8, pp. 853-857, 2007.

[18] W. J. Jiang, S. C. Gao, P. Wittek, and L. Zhao, “Real-time quantifying heart beat rate from facial video recording on a smart phone using kalman filters”, Proceedings of IEEE 16th International Conference on e-Health Networking, Applications and Services, pp. 393-396, 2014.

[19] D. E. King, “Dlib-ml: a machine learning toolkit”, Journal of Machine Learning Research, Vol. 10, pp. 1755-1758, 2009.

[20] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees”, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014. DOI: 10.1109/CVPR.2014.241

[21] R. L. Longini and R. Zdrojkowski, “A note on the theory of backscattering of light by living tissue”, IEEE Transactions on Biomedical Engineering, Vol. 15, No. 1, pp. 4-10, 1968.

[22] V. Vezhnevets, V. Sazonov, and A. Andreeva, “A survey on pixel-based skin color detection techniques”, Proceedings of the International Conference Graphicon, pp.85-92, 2003.

[23] J. Määttä, A. Hadid, and M. Pietikäinen, “Face spoofing detection from single images using micro-texture analysis”, Proc. IEEE Int. Joint Conf. on Biometrics (IJCB), pp.1-7, 2011.

[24] Y. Li, L. M. Po, X. Xu, L. Feng, and F. Yuan, “Face liveness detection and recognition using shearlet based feature descriptors”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp.874-877, 2016.

[25] W. Bao, H. Li, N. Li, and W. Jiang, “A liveness detection method for face recognition based on optical flow field”, Proc. Int. Conf. on Image Anal. Signal Process. (IASP), pp.233-236, 2009.

[26] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee, “Face liveness detection using variable focusing”, Proc. Int. Conf. on Biometrics (ICB), pp.1-6, 2013.

[27] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Z. Li, “Face liveness detection by exploring multiple scenic clues”, Proc. Int. Conf. on Control Automation Robotics & Vision (ICARCV), pp.188-193, 2012.

[28] L. Mei, D. Yang, Z. Feng, and J. Lai, “WLD-TOP based algorithm against face spoofing attacks”, in Chinese Conf. on Biometric, Springer, Cham, pp.135-142, 2015.

[29] H. K. Jee, S. U. Jung, and J. H. Yoo, “Liveness detection for embedded face recognition system”, World Acad. Sci. Eng. Technol., Vol.2, No.6, pp.941-943, 2008.

[30] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes, and S. Sridharan, “Liveness detection based on 3D face shape analysis”, Int. Workshop on Biometrics and Forensics (IWBF), pp.1-4, 2013.

[31] L. Sun, W. B. Huang, and M. H. Wu, “TIR/VIS correlation for liveness detection in face recognition” Proc. 14th Int. Conf. on Comput. Anal. Of Images and Pattern, pp.114-121, 2011.

[32] X. Sun, L. Huang, and C. Liu, “Multispectral face spoofing detection using VIS–NIR imaging correlation”, Int. J. of Wavelets, Multiresolution and Information Processing, Vol.16, No.2, 1840003, 2018.

[33] C. A. Brauer, M. Coca-Perraillon, D. M. Cutler, and A. B. Rosen, “Incidence and mortality of hip fractures in the United States”, JAMA, 302, 1573-1579, 2009.

[34] O. Guzon-Illescas, Elia Perez Fernandez, Natalia Crespí Villarias, Francisco Javier Quirós Donate, Marina Peña, Carlos Alonso-Blas, Alberto García-Vadillo, and Ramon Mazzucchelli, “Mortality after osteoporotic hip fracture: incidence, trends and risk factors”, J. Orthop. Surg. Res., Vol. 14, No. 203, 2019.

[35] D. P. O’Brien, F. A. Luchette, S. J. Pereira, E. Lim, C. S. Seeskin, L. James, S. Miller, K. Davis Jr., J. M. Hurst, J. A. Johannigman, and S. B. Frame, “Pelvic fracture in the elderly is associated with increased mortality”, Surgery, Vol. 132, pp. 710-715, 2002.

[36] T. A. Dechert, T. M. Duane, B. P. Frykberg, M. B. Aboutanos, A. K. Malhotra, and R. R. Ivatury, “Elderly patients with pelvic fracture: interventions and outcomes”, Am. Surg. Vol. 75, pp. 291-295, 2009.

[37] R. Marks, “Hip fracture epidemiological trends, outcomes, and risk factors, 1970-2009”, Int. J. Gen. Med. Vol. 3, pp. 1-17, 2010.

[38] E. M.l Lewiecki, N. C. Wright, J. R. Curtis, E. Siris, R. F. Gagel, K. G. Saag, A. J. Singer, P. M. Steven, and R. A. Adler, “Hip fracture trends in the United States, 2002 to 2015”, Osteoporos Int., Vol. 29, pp. 717-722, 2018.

[39] H. Hagino, N. Endo, A. Harada, J. Iwamoto, T. Mashiba, S. Mori, S. Ohtori, A. Sakai, J. Takada, and T. Yamamoto, “Survey of hip fractures in Japan: recent trends in prevalence and treatment”, J. Orthop. Sci. Vol. 25, pp. 909-914, 2017.

[40] P. Kannus, J. Parkkari, S. Niemi, and H. Sievänen, “Low-trauma pelvic fractures in elderly Finns in 1970-2013”, Calcif Tissue Int. Vol. 97, pp. 577-580, 2015.

[41] S. Andrich, B. Haastert1, E. Neuhaus, K. Neidert, W. Arend, C. Ohmann, J. Grebe, A. Vogt, P. Jungbluth, G. Rösler, J. Windolf, and A. Icks, “Epidemiology of pelvic fractures in Germany: considerably high incidence rates among older people”, PLoS One, Vol. 10, No. 9, e0139078, 2015.

[42] J. P. Grimes, P. M. Gregory, H. Noveck, M. S. Butler, and J. L. Carson, "The effects of time-to-surgery on mortality and morbidity in patients following hip fracture", Am. J. Med. Vol. 112, pp. 702-709, 2002.

[43] P. M. Rommens, C. Arand, A. Hofmann, and D. Wagner, "When and how to operate fragility fractures of the pelvis?", Indian J. Orthop., Vol. 53, pp. 128-137, 2019.

[44] N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell, "A systematic review of fatigue in radiology: is it a problem?", AJR Am. J. Roentgenol., Vol. 210, pp. 799-806, 2018.

[45] T. Shiga, Z. Wajima, and Y. Ohe, "Is operative delay associated with increased mortality of hip fracture patients? Systematic review, meta-analysis, and meta-regression", Can. J. Anaesth., Vol. 55, pp. 146-154, 2008.

[46] M. Falchi, G. A. Rollandi, "CT of pelvic fractures", Eur. J. Radiol., Vol. 50, pp. 96-105, 2004.

[47] A. Pinto, D. Berritto, A. Russo, F. Riccitiello, M. Caruso, M. P. Belfiore, V. R. Papapietro, M. Carotti, F. Pinto, A. Giovagnoni, L. Romano, and R. Grassi, "Traumatic fractures in adults: missed diagnosis on plain radiographs in the emergency department", Acta Biomed., Vol. 89, pp. 111-123, 2018.

[48] E. A. Krupinski, K. S. Berbaum, R. T. Caldwell, K. M. Schartz, and J. Kim, "Long radiology workdays reduce detection and accommodation accuracy", J. Am. Coll. Radiol., Vol. 7, pp. 698-704, 2010.

[49] M. Gribaudoa, P. Piazzolla, F. Porpigliac, E. Vezzetti, and M. G. Violante, "3D augmentation of the surgical video stream: toward a modular approach", Com. Method. Prog. Biomed., Vol. 191, 105505, 2020.

[50] A. S. Chowdhury, J. Burns, B. Sen, A. Mukherjee, J. Yao, and R. M. Summer, "Detection of pelvic fractures using graph cuts and curvatures", 18th IEEE Int. Conf. Image Proc., pp. 1573-1576, 2011.

[51] J. Wu, P. Davuluri, K. R. Ward, C. Cockrell, R. Hobson, and K. Najarian, "Fracture detection in traumatic pelvic CT images", Int. J. Biomed. Imag., 327198, 2012.

[52] R. Lindseya, A. Daluiskia, S. Chopraa, A. Lachapellea, M. Mozera, S. Siculara, D. Hanela, M. Gardnera, A. Guptaa, R. Hotchkissa, and H. Pottera, "Deep neural network improves fracture detection by clinicians", Proc. Natl. Acad. Sci. U S A., Vol. 115, pp. 11591-11596, 2018.

[53] Y. L. Thian, Y. Li, P. Jagmohan, D. Sia, V. E. Y. Chan, and R. T. Tan, "Convolutional neural networks for automated fracture detection and localization on wrist radiographs", Radiology: Artificial Intelligence, Vol. 1, No. 1, 2019.

[54] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, and N. Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network", Skeletal Radiol. Vol. 48, pp. 239-244, 2019.

[55] Y. Sato, Y. Takegami, T. Asamoto, Y. Ono, T. Hidetoshi, R. Goto, A. Kitamura, and S. Honda, "Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures – a multicenter study", BMC Musculoskeletal Disorders, Vol. 22, No. 1, 407, 2020.

[56] C. T. Cheng, C. C. Chen, F. J. Cheng, H. W. Chen, Y. S. Su, C. N. Yeh, I. F. Chung, and C. H. Liao, "A human-algorithm integration system for hip fracture detection on plain radiography: system development and validation study", JMIR Med. Inform., Vol. 8, No. 11, e19416, 2020.

[57] L. Tanzi, E. Vezzetti, R. Moreno, A. Aprato, A. Audisio, and A. Massè, "A. hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach", Eur. J. Radiol., Vol. 133, 109373, 2020.

[58] A. Bar, L. Wolf, O. B. Amitai, E. Toledano, and E. Elnekave, "Compression fractures detection on CT", Proc. SPIE 10134, Med. Imag. 2017: Comp. Aided Diag., 1013440, 2017.

[59] H. R. Roth, Y. Wang, J. Yao, L. Lu, J. E. Burns, and R. M. Summers, "Deep convolutional networks for automated detection of posterior-element fractures on spine CT", Proc. SPIE 9785, Med. Imag 2016: Comp. Aided Diag., 97850P, 2016.

[60] Q. Q. Zhou, J. Wang, W. Tang, Z. C. Hu, Z. Y. Xia, X. S. Li, R. Zhang, X. Yin, B. Zhang, and H. Zhang, "Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: accuracy and feasibility", Kor. J. Radiol., Vol. 21, pp. 869-879, 2020.

[61] C. Rao and Y. Liu, "Three-dimensional convolutional neural network (3D-CNN) for heterogeneous material homogenization", Compt. Mat. Sci., Vol. 184, 109850, 2020.

[62] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations", Med. Image Comput. Comput. Assist. Interv., Vol. 17, pp. 520-527, 2014.

[63] T. Nakao, S. Hanaoka, Y. Nomura, I. Sato, M. Nemoto, S. Miki, E. Maeda, T. Yoshikawa, N. Hayashi, and O. Abe, "Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography", J. Magn. Reson. Imaging., Vol. 47, pp. 948-953, 2018.

[64] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement", arXiv:1804.02767, 2018.

[65] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge", Int. J. Comput. Vis., Vol. 88, pp. 303-338, 2010.

List of Publications

Chapter 2

鵜飼 和歳, Rahman Rashedur, 小橋 昌司:「多点計測投票法による時系列顔画像からの短時間 R-R 間隔推定」, システム制御情報学会論文誌, Vol.31, No.12, pp.403-411, 2018.

Kazutoshi Ukai, Rashedur Rahman, Syoji Kobashi: "Short-time estimation of R-R interval from facial video image with a multiple-measurement-points-voting-method", Transactions of the Institute of Systems, Control and Information Engineers, Vol.31, No.12, pp.403-411, 2018. (in Japanese)

Chapter 3

鵜飼 和歳, Rahman Rashedur, 小橋 昌司:「動画像を用いた顔皮膚血流変化に基づく生体検知法」, 電気学会論文誌 E, Vol.139, No.2, pp.29-37, 2019.

Kazutoshi Ukai, Rashedur Rahman, Syoji Kobashi: "Facial skin blood perfusion change based liveness detection using video images", IEEJ Transactions on Sensors and Micromachines, Vol.139, No.2, pp.29-37, 2019. (in Japanese)

Chapter 4

Kazutoshi Ukai, Rashedur Rahman, Naomi Yagi, Keigo Hayashi, Akihiro Maruo, Hirotsugu Muratsu, Syoji Kobashi: "Detecting pelvic fracture on 3D-CT using deep convolutional neural networks with multi-orientated slab images", Scientific Reports, Vol.11, No.11716, 2021.

Awards

産業技術賞, システム制御情報学会,「多点計測投票法による時系列顔画像からの短時間 R-R 間隔推定」, 鵜飼 和歳, Rahman Rashedur, 小橋昌司, 2019

Industrial Technology Award, The Institute of Systems, Control and Information Engineers, "Short-time estimation of R-R interval from facial video image with a multiple-measurement-points-voting-method", Kazutoshi Ukai, Rashedur Rahman, Syoji Kobashi, 2019. (in Japanese)