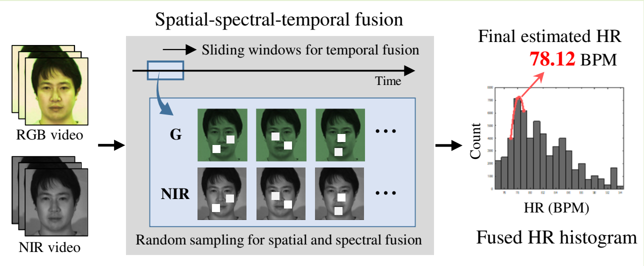
Spatial-Spectral-Temporal Fusion for Remote Heart Rate Estimation

Shiika Kado, Yusuke Monno, *Member, IEEE,* Kazunori Yoshizaki, Masayuki Tanaka, *Member, IEEE,* and Masatoshi Okutomi, *Member, IEEE*

Abstract*—In this paper, we propose a novel heart rate (HR) estimation method using simultaneously recorded RGB and near-infrared (NIR) face videos to improve the robustness of camera-based remote HR estimation against illumination fluctuations and head motions. The key to robust HR estimation is constructing the histogram of HRs for a considered time window by voting candidate HRs that are estimated using different spatial face patches, spectral modalities (i.e., RGB and NIR), and temporal short-time sub-windows. The histogram voting is performed only for the candidate HRs that pass through a reliability check of HR estimation. The final HR estimate for the considered time window is then obtained by detecting the most frequently voted HR bin and performing parabola fitting using its neighboring bins. By spatially, spectrally, and temporally fusing the candidate HRs for majority voting, our method can automatically exploit suitable video sub-regions less affected by illumination fluctuations and head motions to enable robust HR estimation. Through the experiments on 168 RGB-NIR video recordings, we demonstrate that our fusion-based method achieves improved HR estimation accuracy compared with existing methods.


Spatial-spectral-temporal fusion
Sliding windows for temporal fusion
Time
RGB video
NIR video
G
NIR
Random sampling for spatial and spectral fusion
Final estimated HR
78.12 BPM
Count
HR (BPM)
Fused HR histogram

***Index Terms*—Remote heart rate measurement, imaging photoplethysmography, RGB and near-infrared cameras.**

I. Introduction

HEART rate (HR) is one of the most essential vital signs, which provides the physiological and emotional state of a person. HR is typically measured using a photoplethysmography (PPG) sensor attached to human skin. An optical PPG sensor measures light reflected from or transmitted through the skin. Since temporal light intensity change on the skin is caused by blood volume change due to heartbeats, HR can be estimated from the PPG signal [1], [2].

To monitor HR activities for a longer duration, HR measurement using a wearable PPG sensor has been actively studied in recent years [3]. While wearable devices, such as wrist-PPG sensors, are beneficial in that they do not interfere with the daily activities of a user, they have a limitation that the sensors must contact with skin. This requirement is not desirable for people with sensitive skin (e.g., neonates, elderly people, and skin damaged patients) and also may reduce the user's comfort such as during sleep. Therefore, many non-contact HR measurement techniques using various types of sensors (e.g., a Doppler radar [4], [5], a microwave sensor [6], an ultrasound system [7], and a digital camera [8]–[10]) have also been proposed (see [11] for a review). These non-contact-based systems are useful for remote HR measurement applications such as neonate monitoring and telemedicine. While the non-contact systems based on waveform measurement (e.g., radar, microwave, and ultrasound) assume that the user is static, camera-based systems are less constrained and do not require that the user is strictly static. Furthermore, since cameras are now widely spread in our lives, we focus on a camera-based system for remote HR measurement in this study.

With the same principle as the contact PPG sensor, HR can be estimated remotely using a digital camera by recording a face video and detecting subtle pixel intensity change on the face skin [8]–[10]. Camera-based HR measurement allows for various remote vital sensing applications such as monitoring of neonates [12], intraoperative patients [13], or drivers [14] and telemedicine [15]. While camera-based remote HR measurement has progressed greatly in recent years [16]–[19], accurate HR estimation in uncontrolled situations remains difficult due to various types of noise such as ambient light fluctuations and person's head motions. The PPG signal extracted from a face video is susceptible to such noise, which is a main challenge of camera-based remote HR estimation.

To tackle this challenge, in this paper, we propose a novel spatial-spectral-temporal fusion method for robust HR estima-

Manuscript received XXXXX XX, 20XX; revised XXXXX XX, 20XX; accepted XXXXX XX, 20XX. Date of publication XXXXX XX, 20XX; date of current version XXXXX XX, 20XX. This work was partly supported by the MIC/SCOPE #141203024. The associate editor coordinating the review of this paper and approving it for publication was XXXXX XXXXX. (Corresponding author: Yusuke Monno.)

S. Kado, Y. Monno, M. Tanaka, and M. Okutomi are with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan (e-mail: ymonno@ok.sc.e.titech.ac.jp).

K. Yoshizaki is with Olympus Corporation, Hachioji, Tokyo 192-8512, Japan.

Digital Object Identifier 10.1109/JSEN.20XX.XXXXXXX

tion. Different from most existing studies that solely use either an RGB or near-infrared (NIR) domain, we exploit an RGB-NIR face video, which consists of simultaneously recorded and spatially aligned RGB and NIR face videos.

One great advantage of using the NIR domain is the ability to increase the light intensity without disturbing human perception by exploiting an invisible NIR light source. This ability enables HR estimation under low-light or dark conditions. It also can alleviate the effect of ambient light fluctuations, which are more likely to occur in the visible domain, since typical artificial light sources, such as TV and PC monitors, have the spectral power only in the visible wavelengths.

While NIR camera-based HR estimation has been actively studied [20]–[24], it is know that the PPG signal in the NIR domain is weaker than that in the visible domain (especially green-channel wavelengths [25], [26]). Thus, complementally using the RGB and the NIR domains could further improve the robustness of HR estimation under various illumination conditions. However, it raises a new challenge of determining which spectral modality is more suitable for each local face region depending on each illumination condition.

To address this challenge, we propose a novel spatial and spectral face patch sampling and fusion manner that can automatically select suitable local face patches and spectral modalities. Furthermore, to improve the robustness to head motions, we combine the spatial and spectral fusion manner with a novel temporal fusion manner, where we aim at exploiting short-time periods with relatively little head motions.

In our fusion-based method, we estimate HR by taking the RGB-NIR face video of a certain time window as an input. Inspired by the Lam and Kuno method [27], we construct the histogram of candidate HRs that are estimated using different spatial face patches, spectral modalities (i.e., RGB and NIR), and temporal short-time sub-windows. The histogram voting is performed only for the candidate HRs that pass through a reliability check of HR estimation. The final HR estimate for the considered time window is then obtained by detecting the most frequently voted HR bin and performing parabola fitting using its neighboring bins.

The key idea of our method is to collect a lot of measurements using spatial, spectral, and temporal sampling, and to estimate HR by selecting reliable measurements (less affected by illumination fluctuations and head motions) and fusing the selected ones. Our spatial-spectral-temporal sampling and fusion approach significantly increases the possibility of extracting stable video regions even under illumination fluctuations and head motions, contributing to the significant robustness and accuracy improvement of the HR estimation.

In experiments, we evaluate our fusion-based method using RGB-NIR face videos captured by a dual-CCD RGB-NIR camera. Through the evaluation of 168 RGB-NIR video recordings under various illumination conditions and including head motions, we demonstrate that our method achieves improved HR estimation accuracy in comparison with existing methods. We also present the first feasibility evaluation results using a novel single-sensor RGB-NIR camera prototype [28] toward low-cost and compact realization of RGB-NIR camera-based HR estimation. Although we have used the RGB-NIR camera setup in this study, our method could extensible to any multi-spectral/modal camera setups such as in [29]–[32].

This paper is an extended version of our previous conference paper [33]. In this extended study, we have improved the robustness of our method against head motions by proposing an extended method incorporating the temporal fusion. According to this, we have conducted extended experiments using additional RGB-NIR videos including head motions. We also have added the first HR estimation results using a novel single-sensor RGB-NIR camera prototype.

The rest of this paper is organized as follows. Section II briefly reviews related work. Section III details our proposed HR estimation method based on the spatial-spectral-temporal fusion. Section IV shows experimental results in comparison with existing methods. Section V concludes the paper.

II. Related Work

Camera-based remote HR estimation has received increasing attention after the Poh et al. work demonstrates successful HR measurement using a consumer-grade RGB camera [9], [10]. The Poh et al. method extracts temporal RGB intensity traces on a face region of interest (ROI) and then performs independent component analysis (ICA) to extract the PPG signal from the RGB traces. It then applies Fourier transform to the extracted PPG signal to find the most dominant frequency that is assumed to be the heartbeats frequency. Many improved methods have been proposed based on the Poh et al. framework (see the papers [2], [16]–[19] for comprehensive reviews). For example, some methods reduce noise artifacts by mixing the RGB traces to extract the PPG signal [34]–[36], instead of applying ICA. In what follows, we briefly introduce existing methods closely related to our work by focusing on the three aspects: (i) the use of multiple face ROIs, (ii) the removal of noisy time periods, and (iii) the use of non-RGB information.

Some studies have presented a method using multiple local face ROIs to make HR estimation more robust to illumination variations at different face regions [37]–[39]. These methods divide the entire face into local face patches using the detected facial landmarks and perform quality-based fusion or selection of temporal intensity traces extracted from each patch. As a state-of-the-art approach, Lam and Kuno proposed a method that randomly and repeatedly selects face patch pairs and uses corresponding green-channel intensity trace pairs as the inputs of ICA to extract the PPG signal for candidate HR estimation [27]. The final HR estimate is then determined based on the majority voting of candidate HRs from random face patch pairs. Although those methods based on multiple face ROIs have shown improved robustness to illumination variations, their application is still limited to the scenes with relatively little head motions.

To address larger head motions, some studies have attempted to discard the time periods with significant noise components. In [40], [41], quality-based intensity trace evaluation is performed to adaptively remove noisy time periods. In [42], [43], an optimization or a deep learning approach is applied to a constructed spatial-temporal intensity trace matrix to estimate HR by exploiting reliable spatial-temporal regions.

However, when using only an RGB video as these studies do, it is essentially difficult to accurately estimate HR under low-light conditions or ambient light fluctuations from artificial visible light sources.

To overcome the limitation of RGB camera-based methods, some studies have exploited non-RGB information. NIR camera [20]–[24] or thermal camera [44], [45] setups have been the most frequently used to enable HR estimation under low-light conditions or visible light fluctuations. Several methods have exploited multi-spectral/modal information using a multi-spectral camera [29] or multiple cameras with different spectral bands/modalities [31], [32]. These non-RGB camera-based methods usually take one of three approaches: (i) solely use a single band [22], [44], [45], (ii) use all obtained bands [20], [21], [23], [31], or (iii) heuristically select the best band set based on experimental results [24], [29], [32]. However, it remains challenging to adaptively select a suitable band set depending on each illumination condition. Very recently, the study [46] has used an RGB-NIR camera setup, which we first exploited for HR estimation in our earlier study [33]. However, the method in [46] requires background estimation, which is another difficult task in cluttered scenes.

Our study is differentiated from the above-mentioned studies in that we propose a general framework that can spatially, spectrally, and temporally exploit suitable video sub-regions for HR estimation. In this paper, our framework is implemented and validated using an RGB-NIR camera setup.

III. Proposed HR Estimation Method

A. Overall framework with temporal fusion

Figure 1 shows the overall framework of our proposed HR estimation method using an RGB-NIR face video. In this study, we consider the problem of estimating HR for a certain time window (30 seconds in our experiments), assuming the situations that HR does not change drastically within that time duration. To improve the robustness to head motions which may occur in the time window, we introduce a novel temporal fusion manner. Our expectation is that, in real HR monitoring situations such as at working places and driver seats, there would be short-time periods with relatively little head motions, even if there exists a large head motion in the time window. With this expectation, the time window is divided into short-time sub-windows (five seconds in our experiments based on the report of [47]) by a sliding window manner (with one second intervals in our experiments). Then, for each sub-window, the histogram of candidate HRs is constructed based on the spatial and spectral face patch sampling-based HR estimation, as illustrated in Fig. 2. The constructed histograms are then fused to form the final histogram. As shown in the right-hand side of Fig. 1, the final HR estimate is obtained based on majority voting and parabola fitting, assuming that the candidate HRs can reliably and consistently be estimated from the short-time periods less affected by the head motions.

B. Histogram construction based on spatial and spectral face patch sampling-based candidate HR estimation

We next detail our spatial and spectral face patch sampling-based HR estimation method for constructing the histogram of candidate HRs for each short-time sub-window (Fig. 2). Our algorithm is inspired by the Lam and Kuno's method [27] that randomly and repeatedly samples two face patches in the G channel. We extend this method by adding the NIR channel, aiming to spatially and spectrally select suitable local face patches for HR estimaiton.

1) Face landmark detection and tracking: Our algorithm first performs face detection and tracking in three videos, i.e., G, NIR, and G+NIR videos. Since the input RGB and NIR videos are spatially aligned, the G+NIR video is generated by the average of the G and the NIR videos. We use the algorithm in [48] (with the implementation by [49], [50]) to detect 68 face landmarks (see Fig. 2(b)) and apply the algorithm in [27] to track the detected face landmarks between image frames. If the landmarks cannot be detected in a frame, the last detected landmarks in previous frames are copied to that frame.

2) Face patch sampling and signal extraction: Our algorithm then extracts a pair of temporal intensity traces based on our spatial and spectral face patch sampling manner, as shown in Fig. 2(c) and 2(d). Our algorithm randomly and repeatedly samples three pairs of face patches, which are referred to as the G-G pair, the G-N pair, and the N-N pair, respectively. For each pair, temporal intensity traces are extracted as follows.

- G-G pair: Two face patches are randomly selected in the G channel. For each patch, the temporal trace of the averaged G intensity is extracted.
- G-N pair: One face patch is randomly selected in the G+NIR channel. For this patch, the temporal traces of the averaged G and NIR intensities are extracted.
- N-N pair: Two face patches are randomly selected in the NIR channel. For each patch, the temporal trace of the averaged NIR intensity is extracted.

3) Face patch pair-based HR estimation: Our algorithm then follows the HR estimation pipeline of [27] to estimate candidate HRs from each pair of traces. First, moving average filter is applied to the extracted intensity traces to reduce noise. Fast ICA [51] is then performed to estimate the PPG signal from the intensity traces. Detrending filter [52] and moving average filter are then applied to the PPG signal to remove trends and reduce noise. Welch's power spectral density calculation [53] is then applied to the PPG signal to find the most dominant frequency between 0.7Hz to 4Hz, which is assumed to be the frequency of the heartbeats. The estimated HR is obtained in the form of beats per minute (BPM) by multiplying the corresponding frequency by 60.

4) Histogram voting and fusion: To use only suitable face patch pairs for the final HR estimation, histogram voting is performed based on the reliability of the estimated HR [27]. The reliability is defined by the ratio of the spectral power of the most dominant frequency to that of the second dominant frequency, which is indicated as v_1/v_2 in Fig. 2(e). This ratio implies how dominant the heartbeats frequency is. Thus, the higher ratio indicates more reliable HR estimation. When the ratio is higher than a threshold value T_r, the estimated HR is round to the integer value and voted to the corresponding histogram bin. The random patch sampling and the histogram voting are repeated K times to construct the histogram of reliably estimated candidate HRs. As the result of the histogram

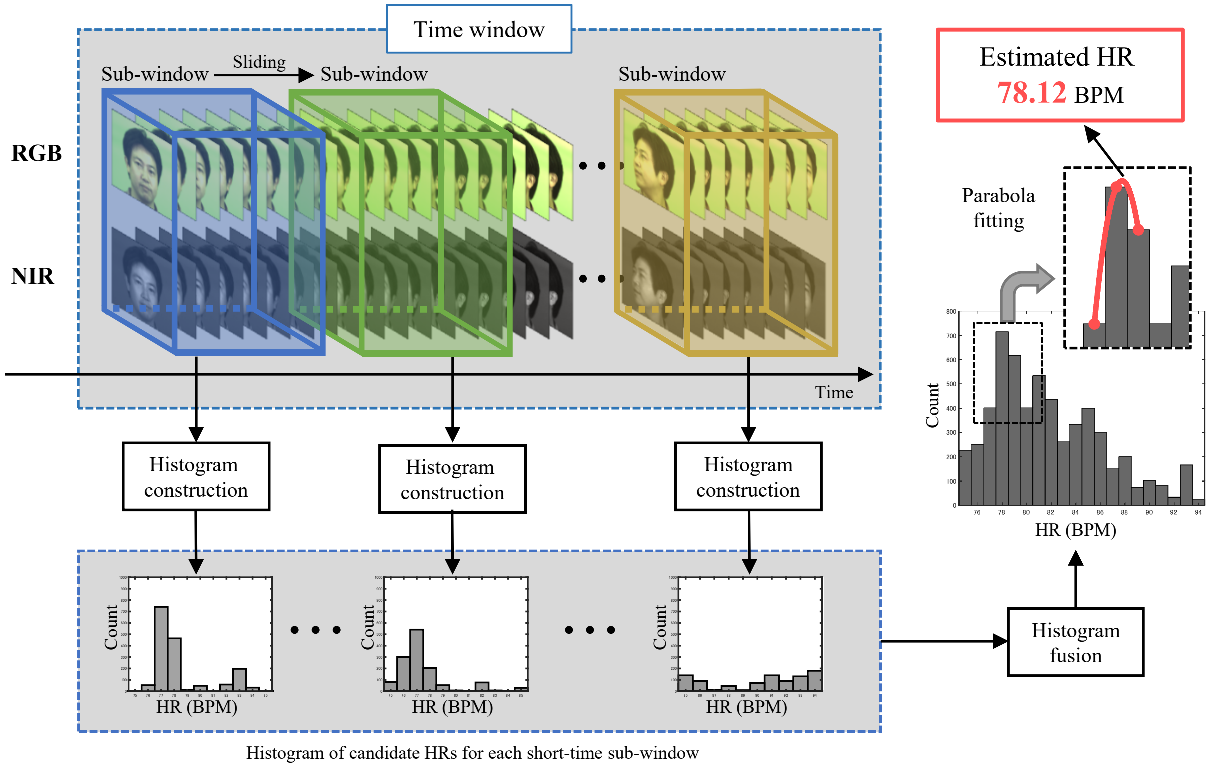


Fig. 1: The overall framework of our proposed HR estimation method using an RGB-NIR face video. We refer to Section III-A for detailed explanation and Fig. 2 for the overall flow of the histogram construction for each short-time sub-windows.

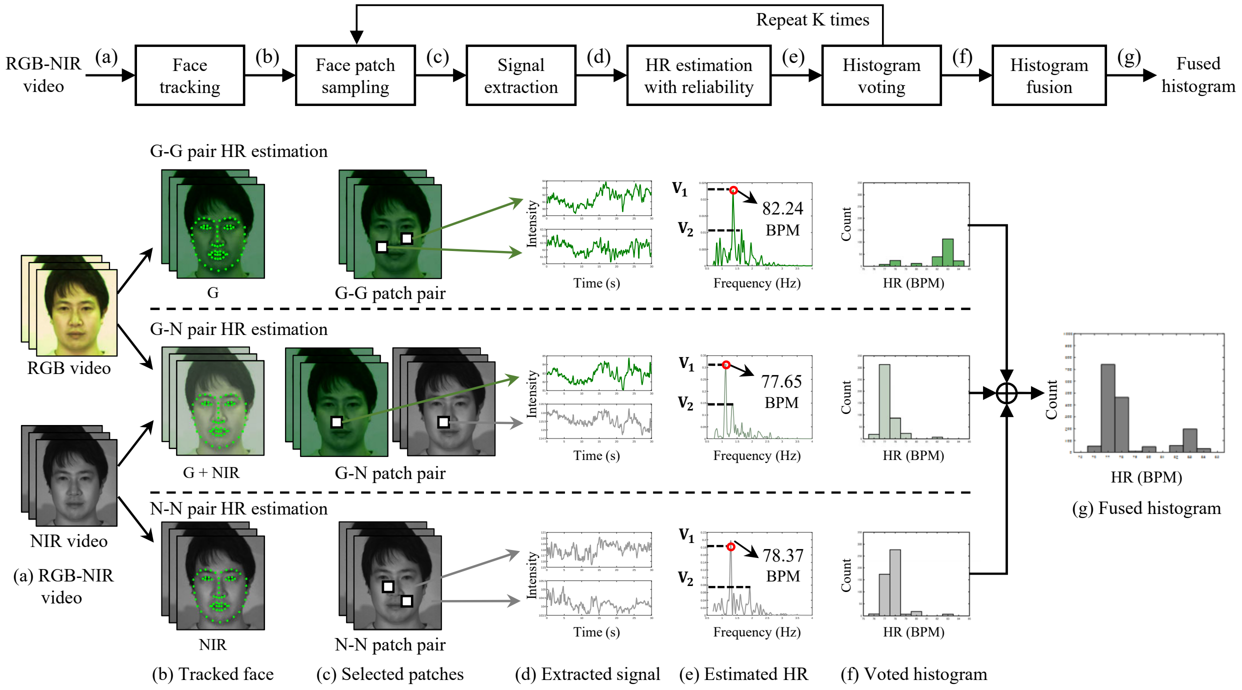


Fig. 2: The overall flow of the histogram construction for each short-time sub-windows based on the spatial and spectral face patch sampling-based candidate HR estimation. We refer to Section III-B for detailed explanation.

voting, we obtain three histograms for the G-G, the G-N, and the N-N pairs, respectively (see Fig. 2(f)). To obtain one histogram for each short-time sub-window, histogram fusion is performed by adding the three histograms.

C. Final HR estimate calculation

To derive the final HR estimate for the considered time window, the constructed histograms for each short-time sub-

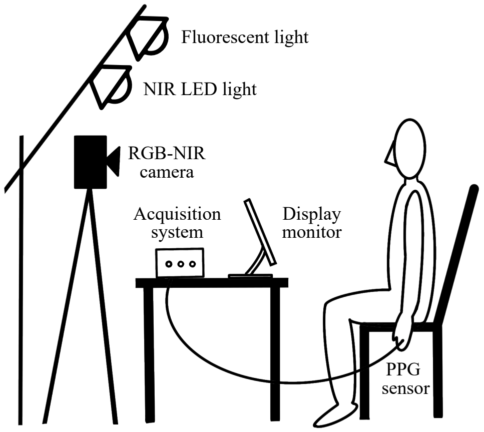
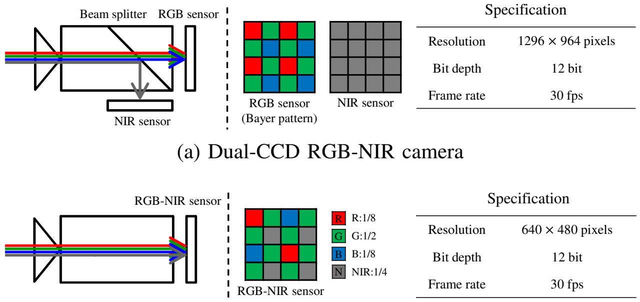


Fig. 3: The experimental setup.



(a) Dual-CCD RGB-NIR camera

(b) Single-sensor RGB-NIR camera

Fig. 4: RGB-NIR camera types.

windows are fused to form the final histogram. In the final histogram, the most frequently voted HR bin is considered as the most reliably and consistently estimated HR exploiting stable video sub-regions less affected by illumination fluctuations and head motions considering the spatial-spectral-temporal domain. The final HR estimate is obtained in the real value precision by performing parabola fitting using that bin and its neighboring bins. The peak of the fitted parabola is the estimated HR for the considered time window.

IV. Experimental Results

A. Setups and data collection

Figure 3 shows the experimental setup for collecting data. We used two types of an RGB-NIR camera, as shown in Fig. 4. The first one is a dual-CCD RGB-NIR camera (AD-130GE, JAI Ltd., Japan), which uses a beam splitter and two image sensors. The second one is a single-sensor RGB-NIR camera prototype [28], which uses a single image sensor equipped with an RGB-NIR filter array. Both cameras can simultaneously record RGB and NIR videos without misalignment.

The subjects were asked to sit in a chair, which was placed at a distance of 1.5 meter from the camera. A contact PPG sensor (Procomp Infinity T7500M, Thought Technology Ltd., Canada) was attached to the subject's finger to acquire reference HR for the evaluation. A display monitor was set on the table in front of the subject to simulate a situation with illumination fluctuations in some scenes. Total 38 subjects (35 Eastern Asians and 3 Southeastern Asians) with both gender (8 females) and different age (20's - 60's) took part in the experiments. The experimental protocols were approved by the research ethics committees of Tokyo Institute of Technology and Olympus Corporation. The informed consent was obtained from all subjects.

TABLE I: Conditions for each scene. The NIR light was turned on for all scenes.

		Fluorescent (FL) light	Illumination fluctuations	Number of videos
Stationary scenes	Scene 1	600 lux	Without	32
	Scene 2	50 lux	Without	32
	Scene 3	50 lux	Movie	32
Motion scenes	Scene 4	600 lux	Without	43
	Scene 5	50 lux	Movie	29

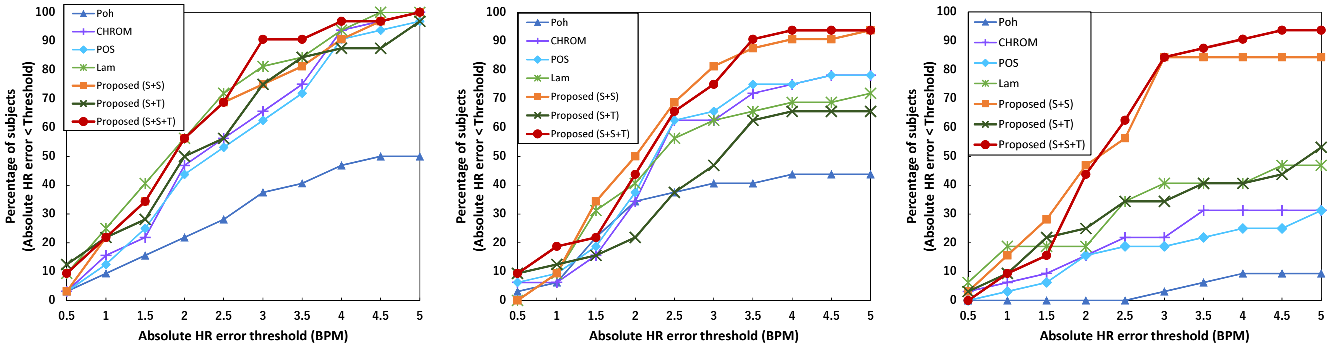
TABLE II: Compared methods.

Methods	Spatial	Spectral	Temporal
Poh [10]			
CHROM [34]			
POS [35]			
Lam [27]	✓		
Proposed (S+S)	✓	✓	
Proposed (S+T)	✓		✓
Proposed (S+S+T)	✓	✓	✓

We conducted the experiments for five scenes to evaluate the robustness of our method against illumination fluctuations and head motions. Table I summarizes the experimental conditions for each scene. The first class of experiments was conducted in the stationary scenes, where the subjects were asked to sit still in the chair. The second class of experiments was conducted in the motion scenes, where the subjects were asked to perform some tasks including head motions. As the lighting setup, we used a fluorescent (FL) light for visible wavelengths and two NIR LEDs, which were placed at the right front and the left front of the subject. The NIR LEDs have a light emission wavelength range from 760nm to 940nm with its peak at 850nm. We also used the display monitor to play a movie with light fluctuations for some scenes. The duration of the video (i.e., the considered time window) is 30 seconds for all scenes. The video data was generated in Motion JPEG 2000 format from the original image frames captured in the RAW data format. The conditions for each scene are further detailed in the result subsections.

B. Compared methods

We compared our method with the Poh et al. method [10] as a baseline method and the CHROM method [34], the POS method [35], and the Lam and Kuno's method [27] as state-of-the-art methods. Table II shows the property of each method. The Poh's method, the CHROM method, and the POS method use a fixed face ROI and RGB channels. The Poh's method applies ICA to the RGB traces to extract the PPG signal. The CHROM method uses a chrominance-based approach and derives two orthogonal color difference signals



(a) Scene 1: FL light (600 lux) + NIR light (b) Scene 2: FL light (50 lux) + NIR light (c) Scene 3: FL light (50 lux) + NIR light + Movie

Fig. 5: The comparison of HR estimation accuracy for the stationary scenes under three illumination conditions.

from the RGB traces. Assuming the standardized skin color, the PPG signal is computed as a linear combination of the two chrominance signals. The POS method derives the PPG signal as a linear combination of the RGB traces using a plane orthogonal to the skin tone. We used a manually selected left and right cheek regions as a fixed face ROI for the Poh's, the CHROM, and the POS methods. We employed iPhys toolbox by McDuff and Blackford [54] to implement the CHROM and the POS methods. The Lam's method uses randomly sampled two face patches and the G channel. In other words, the Lam's method employs the spatial face ROI sampling and fusion. Our proposed method performs the spatial, spectral and temporal face ROI fusion using the G and the NIR channels (noted as proposed (S+S+T)). To assess the effectiveness of each fusion manner, we also compared the method using only the spatial and spectral fusion (noted as proposed (S+S)) and the method using only the spatial and temporal fusion with the G channel (noted as proposed (S+T)). For the Lam's method and the three proposed methods, we used the same parameter values, $K = 500$ and $T_r = 2$, for the random patch sampling and the reliability check. The size of each face patch is set as 30% of the horizontal length of the detected face. We applied the same parameter values as the original method proposed by Lam and Kuno [27], which were empirically determined in their study. For the two proposed methods with the temporal fusion, we used five seconds short-time sub-windows with one seconds intervals. To focus on the differences in used fusion manners for each method, we only have changed the way of taking face ROIs to obtain the temporal intensity traces. All the other processes, such as the face tracking and the PPG signal extraction, are the same for all methods.

To compute reference HR using the contact PPG sensor, we first detected the peaks of the contact sensor's PPG signal. We then manually confirmed that the detected peaks are correct and the contact sensor's PPG signal is reliable as a reference. Then, we calculated inter-beat time intervals (IBIs) between every two successive peaks. The average of all IBIs in the considered time window was then calculated to derive the average IBI for that time window. Finally, the contact PPG sensor's HR for that time window was computed by dividing 60 by the average IBI.

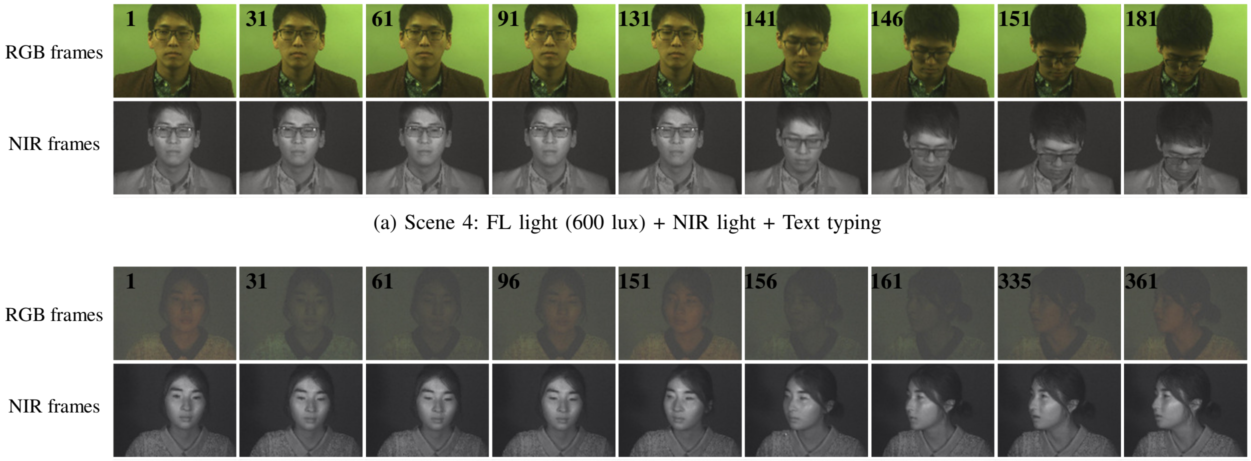
C. *Results for stationary scenes*

We first evaluate the HR estimation accuracy for the stationary scenes under three different illumination conditions. These experiments were performed by using the dual-CCD camera. Figure 5 shows the HR estimation accuracy in comparison with the contact PPG sensor. In the result graphs, the horizontal axis shows the absolute HR estimation error threshold in BPM. The vertical axis shows the percentage of subjects whose absolute HR estimation error is less than the threshold of the horizontal axis. In each graph, a more upper-left line indicates that the method provides better performance and can estimate HR with fewer absolute errors for more subjects.

1) Scene 1: The videos were recorded under the FL light (600 lux) and the NIR light. Because this condition provides sufficient lighting for both the RGB and the NIR videos, the Lam's method and our three methods provide similar results. These methods incorporate the spatial face ROI sampling and fusion manner, contributing to slightly better performance than the CHROM and the POS methods, which are state-of-the-art methods using a fixed face ROI.

2) Scene 2: The videos were recorded under the FL light (50 lux) and the NIR light. This is a low-light condition, under which the RGB camera-based methods (Poh, CHROM, POS, Lam and proposed (S+T)) are difficult to accurately estimate HR, though the CHROM and the POS methods show much better performance than the baseline Poh's method by effectively suppressing noise artifacts with color channel mixing. In contrast, our methods with the spectral fusion using both the G and the NIR channels (proposed (S+S) and proposed (S+S+T)) can successfully improve the robustness of HR estimation by spatially and spectrally selecting suitable face patch pairs from the G-G, the G-N and the N-N pairs, while the Lam's method only applies the spatial fusion using the G-G pairs.

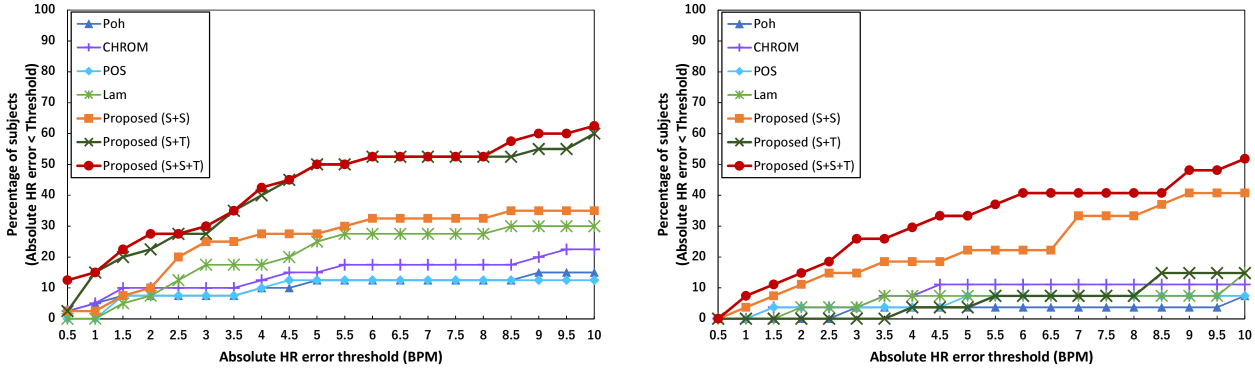
3) Scene 3: The videos were recorded under the FL light (50 lux) and the NIR light. A movie had also been played on the display monitor. This is a challenging condition for the RGB camera-based methods because the videos were recorded under both a low-light condition and light fluctuations in the visible domain. Compared with Scene 2, the accuracy of the RGB camera-based methods (Poh, CHROM, POS, Lam and proposed (S+T)) decreases due to the light fluctuations from



(a) Scene 4: FL light (600 lux) + NIR light + Text typing

(b) Scene 5: FL light (50 lux) + NIR light + Movie + Speaking

Fig. 6: The example image frames for the motion scenes.



(a) Scene 4: FL light (600 lux) + NIR light + Motion

(b) Scene 5: FL light (50 lux) + NIR light + Movie + Motion

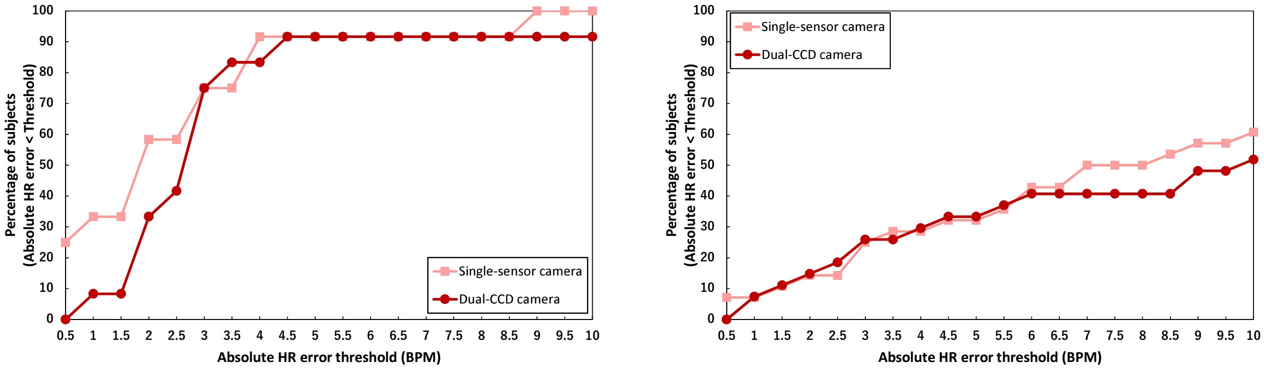
Fig. 7: The comparison of HR estimation accuracy for the motion scenes.

the monitor. For this challenging condition, the CHROM and the POS methods provide limited performance because these methods fail to sufficiently suppress the artifacts derived from illumination fluctuations. Although the Lam's method shows better performance than the other existing methods, it only provides roughly 47% success ratio for the error threshold of 5 BPM. In contrast, our methods with the spectral fusion (proposed (S+S) and proposed (S+S+T)) remain stable and present better HR estimation accuracy than the other methods.

D. Results for motion scenes

We next evaluate the HR estimation accuracy for the motion scenes. These experiments were also performed by using the dual-CCD camera. Two illumination conditions were tested for the motion scenes, which are detailed later. In the motion scenes, for each video recording, the subjects were asked to sit in the chair and perform one of the three tasks with head motions. The details of motion protocols for each task are as follows. (i) Tracking the ball on the display monitor: The subject was asked to track the ball displayed on the monitor, which entails about 90-degrees vertical head rotation at a speed of about 18 degrees per second. The task was continued during video recording. (ii) Typing displayed texts on a smartphone: The subject was asked to look at the display monitor and memorize the displayed texts for five seconds. In the next five seconds, the subject was asked to type the texts on a smartphone, which entails look-down and look-up head motions. The task was repeated during video recording. (iii) Speaking with a person who is beside the subject: The subject was asked to face forward for five seconds. In the next five seconds, the subject was asked to speak with the person beside the subject, which entails about 90-degrees horizontal head rotation and facial movements associated with speaking. The task was repeated during video recording.

1) Scene 4: The videos were recorded under the FL light (600 lux) and the NIR light, which is a sufficient condition for both the RGB and the NIR videos. Figure 6(a) shows example image frames for this illumination condition with the text typing task. Figure 7(a) shows the result for 43 video recordings (15 videos for the ball tracking task, 15 videos for the text typing task, and 13 videos for the speaking task). Generally, the head motions make the HR estimation very difficult and all methods present worse HR estimation accuracy compared with the stationary scenes. In the motion scenes, the methods without the temporal fusion (Poh, CHROM, POS, Lam, and proposed (S+S)) are not able to estimate HR accurately, because the temporal PPG signal extracted from the whole time window is very susceptible to motion noise. For the



(a) Scene 3: FL light (50 lux) + NIR light + Movie

(b) Scene 5: FL light (50 lux) + NIR light + Movie + Motion

Fig. 8: The evaluation of HR estimation accuracy when using the single-sensor RGB-NIR camera.

situation with large head motions, sufficiently suppressing the motion artifacts using the whole time-window RGB signals is difficult, even with state-of-the-art color channel mixing techniques of the CHROM and the POH methods. In contrast, our methods with the temporal fusion (proposed (S+T) and proposed (S+S+T)) can improve the accuracy of HR estimation. These results validate the effectiveness of the temporal fusion that tries to exploit short-time periods less affected by the head motions.

2) Scene 5: The videos were recorded under the FL light (50 lux) and the NIR light. A movie had also been played on the display monitor. Figure 6(b) shows example image frames for this illumination condition with the speaking task. This scene is very challenging because the videos include all considered noise sources, i.e., low lightness, light fluctuations, and head motions. This condition assumes the HR monitoring situations such that a person watches a movie in a theater or drives a car in the night. Figure 7(b) shows the result for 29 video recordings (16 videos for the ball tracking task and 13 videos for the speaking task). The RGB camera-based methods (Poh, CHROM, POS, Lam, proposed (S+T)) fails to estimate HR in this scene because of low lightness and light fluctuations in the visible domain. On the other hand, our method with the spectral fusion (proposed (S+S)) shows better performance owing to the use of the NIR video. In addition, our method with all the spatial, spectral, and temporal fusion manners, (proposed (S+S+T)) can further improve the robustness of HR estimation for this very challenging condition by exploiting suitable video sub-region less affected by both the light fluctuations and the head motions in the spatial-spectral-temporal domain.

Table III shows the overall comparison of HR estimation accuracy and compares the percentage of subjects whose absolute HR estimation error is less than 5 BPM for stationary scenes and less than 10 BPM for more challenging motion scenes. We can confirm that our proposed method based on the spatial-spectral-temporal fusion (proposed (S+S+T)) achieves the best performance for both stationary and motion scenes.

Although our method clearly outperforms the existing methods as summarized in Table III, it still has some limitations. (i) Our method is based on the expectation that HR can be derived reliably from some of the short-time sub-windows. Therefore, our method may fail to accurately estimate HR under the situations that continuous and hard motions exist during a whole time window, such as fitness scenes. (ii) Our method derives the final HR estimate by the majority voting of candidate HRs from different short-time sub-windows. This means that the most dominant and consistent HR in that time window is considered as our method's output. Therefore, if a large HR spread exists during the considered time window, our method may fail to sufficiently track that HR spread. Although the use of a shorter time window may increase the tracking capability to the HR spread, it may decrease the robustness of HR estimation as a trade-off.

TABLE III: The overall comparison of HR estimation accuracy. The table compares the percentage of subjects whose absolute HR estimation error is less than 5 BPM for stationary scenes and less than 10 BPM for challenging motion scenes.

	Percentage of subjects (Error $<$ 5 BPM)			Percentage of subjects (Error $<$ 10 BPM)	
	Stationary scenes			Motion scenes	
	Scene1	Scene2	Scene3	Scene4	Scene5
Poh	50.00	43.75	9.38	15.00	7.41
CHROM	96.88	78.13	31.25	12.50	7.41
POS	**100.00**	78.13	31.25	22.50	11.11
Lam	**100.00**	71.88	46.88	30.00	14.81
Proposed (S+S)	**100.00**	**93.75**	84.38	35.00	40.74
Proposed (S+T)	96.88	65.63	53.13	60.00	14.81
Proposed (S+S+T)	**100.00**	**93.75**	**93.75**	**62.50**	**51.85**

E. Results for the single-sensor RGB-NIR camera

Compared with the dual-CCD RGB-NIR camera, the single-sensor RGB-NIR camera enables lower cost and more compact video recordings as current consumer RGB cameras do, and thus it is more suitable for remote HR estimation applications in various embedded systems. However, the single-sensor camera is still in the research and development stage, we here evaluate its feasibility for remote HR estimation.

For the evaluation, we captured the same scenes by using the dual-CCD and the single-sensor cameras in parallel. Figure 8(a) and 8(b) respectively show the result for 19 video recordings under the same condition as Scene 3 and the result for 29 video recordings under the same condition as Scene 5. From these results, we can confirm that the single-sensor camera presents similar trends with the dual-CCD camera, though there are slight performance differences, which could be due to the different sensor characteristics such as in the pixel size and the camera spectral sensitivity. This feasibility evaluation demonstrates the potential of the single-sensor RGB-NIR camera, as well as the dual-CCD RGB-NIR camera, for remote HR estimation.

V. Conclusion

In this paper, we have proposed a novel spatial-spectral-temporal fusion method that can significantly improve the robustness of camera-based remote HR estimation against illumination fluctuations and head motions. Our method takes an RGB-NIR face video as an input and automatically exploits stable video sub-regions less affected by illumination fluctuations and head motions by fusing the candidate HRs estimated using different face patches, spectral modalities, and temporal short-time sub-windows. We have validated the effectiveness of our fusion-based method using two types of an RGB-NIR camera. The experimental comparison with existing methods has demonstrated that our method can achieve improved HR estimation accuracy for challenging scenes including illumination fluctuations and head motions. In future work, we will investigate the extension of our method to estimate other cardiac parameters such as inter-beat intervals [55] and heart rate variability [56].

References

[1] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, 2007.

[2] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 3, pp. 463–477, 2016.

[3] D. Biswas, N. Simões-Capela, C. V. Hoof, and N. V. Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6560–6570, 2019.

[4] W. Massagram, V. M. Lubecke, A. HØst-Madsen, and O. Boric-Lubecke, "Assessment of heart rate variability and respiratory sinus arrhythmia via Doppler radar," *IEEE Trans. on Microwave Theory and Techniques*, vol. 57, no. 10, pp. 2542–2549, 2009.

[5] D. Obeid, S. Sadek, G. Zaharia, and G. E. Zein, "Multitunable microwave system for touchless heartbeat detection and heart rate variability extraction," *Microwave and Optical Technology Letters*, vol. 52, no. 1, pp. 192–198, 2010.

[6] G. Lu, F. Yang, Y. Tian, X. Jing, and J. Wang, "Contact-free measurement of heart rate variability via a microwave sensor," *Sensors*, vol. 9, pp. 9572–9581, 2009.

[7] M. Ambrosanio, S. Franceschini, G. Grassini, and F. Baselice, "A multichannel ultrasound system for non-contact heart rate monitoring," *IEEE Sensors Journal*, vol. 20, no. 4, pp. 2064–2074, 2020.

[8] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[9] M. Z. Poh, D. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[10] M. Z. Poh, D. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[11] J. Kranjec, S. Beguš, J. Drnovšek, and G. Geršak, "Novel methods for noncontact heart rate measurement: A feasibility study," *IEEE Trans. on Instrumentation and Measurement*, vol. 63, no. 4, pp. 838–847, 2014.

[12] L. A. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. B. Oetomo, and W. Verkruysse, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit – A pilot study," *Early Human Development*, vol. 89, no. 12, pp. 943–948, 2013.

[13] A. Trumpp, J. Lohr, D. Wedekind, M. Schmidt, M. Burghardt, A. R. Heller, H. Malberg, and S. Zaunseder, "Camera-based photoplethysmography in an intraoperative setting," *Biomedical Engineering Online*, vol. 17, no. 1, pp. 33–1–19, 2018.

[14] S. Leonhardt, L. Leicht, and D. Teichmann, "Unobtrusive vital sign monitoring in automotive environments – A review," *Sensors*, vol. 18, no. 9, pp. 3080–1–38, 2018.

[15] F. Zhao, M. Li, Y. Qian, and J. Z. Tsien, "Remote measurements of heart and respiration rates for telemedicine," *PloS One*, vol. 8, no. 10, pp. e71 384–1–14, 2013.

[16] A. Sikdar, S. K. Behera, and D. P. Dogra, "Computer-vision-guided human pulse rate estimation: A review," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 91–105, 2016.

[17] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, B. Karasfi, Y. S. Ali, and F. Mériaudeau, "Heart rate estimation using facial video: A review," *Biomedical Signal Processing and Control*, vol. 38, pp. 346–360, 2017.

[18] M. Harford, J. Catherall, S. Gerry, J. D. Young, and P. J. Watkinson, "Availability and performance of image-based, non-contact methods of monitoring heart rate, blood pressure, respiratory rate, and oxygen saturation: A systematic review," *Physiological Measurement*, vol. 40, no. 6, pp. 06TR01–1–17, 2019.

[19] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Trans. on Instrumentation and Measurement*, vol. 68, no. 10, pp. 3600–3615, 2019.

[20] M. van Gastel, S. Stuijk, and G. de Haan, "Motion robust remote-PPG in infrared," *IEEE Trans. on Biomedical Engineering*, vol. 62, no. 5, pp. 1425–1433, 2015.

[21] S. B. Park, G. Kim, H. J. Baek, J. H. Han, and J. H. Kim, "Remote pulse rate measurement from near-infrared videos," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1271–1275, 2018.

[22] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavany, "SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared," *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1353–1362, 2018.

[23] W. Wang, A. C. Den Brinker, and G. de Haan, "Full video pulse extraction," *Biomedical Optics Express*, vol. 9, no. 8, pp. 3898–3914, 2018.

[24] F. Wurtenberger, T. Haist, C. Reichert, A. Faulhaber, T. Bottcher, and A. Herkommer, "Optimum wavelengths in the near infrared for imaging photoplethysmography," *IEEE Trans. on Biomedical Engineering*, 2019.

[25] L. F. Corral, G. Paez, and M. Strojnik, "Optimal wavelength selection for non-contact reflection photoplethysmography," *Proc. of SPIE*, vol. 8011, pp. 801 191–1–7, 2011.

[26] E. B. Blackford, J. R. Estepp, and D. McDuff, "Remote spectral measurements of the blood volume pulse with applications for imaging photoplethysmography," *Proc. of SPIE*, vol. 10501, pp. 105 010Z–1–8, 2018.

[27] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 3640–3648, 2015.

[28] Y. Monno, H. Teranaka, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Single-sensor RGB-NIR imaging: High-quality system design and prototype implementation," *IEEE Sensors Journal*, vol. 19, no. 2, pp. 497–507, 2019.

[29] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Trans. on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, 2014.

[30] Y. Monno, S. Kikuchi, M. Tanaka, and M. Okutomi, "A practical one-shot multispectral imaging system using a single image sensor," *IEEE Trans. on Image Processing*, vol. 24, no. 10, pp. 3048–3059, 2015.

[31] O. Gupta, D. McDuff, and R. Raskar, "Real-time physiological measurement and visualization using a synchronized multi-camera system," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 46–53, 2016.

[32] E. B. Blackford and J. R. Estepp, "A multispectral testbed for cardiovascular sensing using imaging photoplethysmography," *Proc. of SPIE*, vol. 10072, pp. 100 720R–1–13, 2017.

[33] S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Remote heart rate measurement from RGB-NIR video based on spatial and spectral face patch selection," *Proc. of Int. Conf.*

of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5676–5680, 2018.

[34] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[35] W. Wang, B. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote-PPG," *IEEE Trans. on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[36] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, 2014.

[37] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[38] P. Gupta, B. Bhowmick, and A. Pal, "Accurate heart-rate estimation from face videos using quality-based fusion," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, pp. 4132–4136, 2017.

[39] Z. Wang, X. Yang, and K.-T. Cheng, "Accurate face alignment and adaptive patch selection for heart rate estimation from videos under realistic scenarios," *PloS One*, vol. 13, no. 5, pp. e0 197 275–1–25, 2018.

[40] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4321–4328, 2014.

[41] P. Gupta, B. Bhowmik, and A. Pal, "Robust adaptive heart-rate monitoring using face videos," *Proc. of IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp. 530–538, 2018.

[42] S. Tulyakov, X. A. Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Seve, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2396–2404, 2016.

[43] X. Niu, X. Zhao, H. Han, A. Das, S. Shan, and X. Chen, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2019.

[44] M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Trans. on Biomedical Engineering*, vol. 54, no. 8, pp. 1418–1426, 2007.

[45] K. Hamedani, Z. Bahmani, and A. Mohammadian, "Spatio-temporal filtering of thermal video sequences for heart rate estimation," *Expert Systems with Applications*, vol. 54, pp. 88–94, 2016.

[46] K. Kurihara, D. Sugimura, and T. Hamamoto, "Adaptive fusion of RGB/NIR signals based on face/background cross-spectral analysis for heart rate estimation," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, pp. 4534–4538, 2019.

[47] Y.-P. Yu, P. Raveendran, and C.-L. Lim, "Dynamic heart rate measurements from video sequences," *Biomedical Optics Express*, vol. 6, no. 7, pp. 2466–2480, 2015.

[48] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1867–1874, 2014.

[49] D. E. King, "Dlib-ml: A machine learning toolkit," *Jounal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[50] Y. Nirkin, I. Masi, A. T. Trân, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pp. 98–105, 2018.

[51] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[52] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE Trans. on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

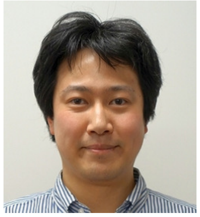
[53] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[54] D. McDuff and E. B. Blackford, "iPhys: An open non-contact imaging-based physiological measurement toolbox," *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6521–6524, 2019.

[55] Y. Maki, Y. Monno, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Inter-beat interval estimation from facial video based on reliability of BVP signals," *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6525–6528, 2019.

[56] Y. Sun, S. Hu, V. Azorin-Peris, R. Kalawsky, and S. Greenwald, "Noncontact imaging photoplethysmography to effectively access pulse rate variability," *Journal of Biomedical Optics*, vol. 18, no. 6, pp. 061 206–1–9, 2013.

Shiika Kado received the B.E., and M.E. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 2018, and 2020, respectively. She has been a network engineer at NTT Communications Corporation. Her research interests is in biomedical engineering. She won the first prize of the paper competition of JPComp2018 organized by IEEE Japan Chapter of Engineering in Medicine and Biology Society in 2018.



Yusuke Monno received the B.E., M.E., and Ph.D degrees from Tokyo Institute of Technology, Tokyo, Japan, in 2010, 2011, and 2014, respectively. He is currently a postdoctoral researcher with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology. From Nov. 2013 to Mar. 2014, he joined the Image and Visual Representation Group at École Polytechnique Fédérale de Lausanne as a research internship student. His research interests are in both theoretical and practical aspects of image processing, computer vision, and biomedical engineering.

Kazunori Yoshizaki received the B.E., and M.E. degrees from Tokyo University of Agriculture and Technology, Tokyo, Japan, in 1998, and 2000, respectively. He was a Research Engineer at Seiko Epson Corporation from 2000 to 2008. Since 2008, he has been a Senior Researcher at Olympus Corporation, where he is engaged in research and development in image processing algorithms and computer vision at the Research and Development Center. His work focuses on biomedical imaging.

Masayuki Tanaka received his bachelor's and master's degrees in control engineering and Ph.D. degree from Tokyo Institute of Technology in 1998, 2000, and 2003, respectively. He was a Software Engineer at Agilent Technology from 2003 to 2004. He was a research Scientist at Tokyo Institute of Technology from 2004 to 2008. He was an Associated Professor at the Graduate School of Science and Engineering, Tokyo Institute of Technology from 2008 to 2016. He was a Visiting Scholar with the Department of Psychology, Stanford University from 2013 to 2014. He was an Associated Professor at the School of Engineering, Tokyo Institute of Technology from 2016 to 2017. He was a Senior Researcher at National Institute of Advanced Industrial Science and Technology from 2017 to 2020. Since 2020, he has been an Associated Professor at School of Engineering, Tokyo Institute of Technology.

Masatoshi Okutomi received the B.Eng. degree from the Department of Mathematical Engineering and Information Physics, the University of Tokyo, Tokyo, Japan, in 1981, and the M.Eng. degree from the Department of Control Engineering, Tokyo Institute of Technology, Tokyo, in 1983. He joined the Canon Research Center, Canon Inc., Tokyo, in 1983. From 1987 to 1990, he was a Visiting Research Scientist with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He received the Dr.Eng. degree from Tokyo Institute of Technology, in 1993, for his research on stereo vision. Since 1994, he has been with Tokyo Institute of Technology, where he is currently a Professor with the Department of Systems and Control Engineering, the School of Engineering.