Non-Contact Heart Rate Estimation via Adaptive RGB/NIR Signal Fusion

Kosuke Kurihara, Daisuke Sugimura, *Member, IEEE*, Takayuki Hamamoto, *Member, IEEE*

***Abstract*—We propose a non-contact heart rate (HR) estimation method that is robust to various situations, such as bright, low-light, and varying illumination scenes. We utilize a camera that records red, green, and blue (RGB) and near-infrared (NIR) information to capture the subtle skin color changes induced by the cardiac pulse of a person. The key novelty of our method is the adaptive fusion of RGB and NIR signals for HR estimation based on the analysis of background illumination variations. RGB signals are suitable indicators for HR estimation in bright scenes. Conversely, NIR signals are more reliable than RGB signals in scenes with more complex illumination, as they can be captured independently of the changes in background illumination. By measuring the correlations between the lights reflected from the background and facial regions, we adaptively utilize RGB and NIR observations for HR estimation. The experiments demonstrate the effectiveness of the proposed method.**

***Index Terms*—Remote vital sensing, RGB/NIR camera, Varying illumination**

I. Introduction

HEART rate (HR) is an essential vital sign based on which the physiological and emotional states of a person can be assessed [1], [2], [3], [4]. Traditional HR estimation methods require contact-type sensors, such as electrocardiograms [5] and pulse oximetry sensors [6]. However, the restrictions associated with these sensors make subjects uncomfortable.

In the last decade, camera-based non-contact HR estimation methods have attracted considerable research attention [7], [8]. Cameras can capture temporal skin color variations arising from changes in the amount of blood circulated during a cardiac cycle [9]; thus, HR can be estimated using the time-series signals recorded in the videos. Notably, effective video-based methods have several potential applications, for example, in human-robot interaction [10], telemedicine [11], and driver monitoring systems [12].

A large number of researchers have developed video-based HR estimation methods using a monocular red, green, and blue (RGB) camera [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. However, their effectiveness is limited in uncontrolled scenes such as low-light scenes or scenes with varying illumination. In low-light scenes, only minimal light can enter the camera. In scenes with varying illuminations, the videos contain dominant color casts caused by varying background illumination. These characteristics affect the accuracy of the measurements of skin color changes attributable to cardiac pulses.

To overcome these problems, researchers have proposed HR estimation methods using near-infrared (NIR) cameras and flash units [12], [27], [28], [29], [30], [31]. Because NIR cameras can capture only the NIR components of incoming light, the influences of varying background illuminations can be suppressed for HR estimation. However, NIR signals yield less accurate HR estimation compared with RGB signals because of the light absorption characteristics of blood. According to previous studies [12], [31], [32], [33], blood tends to absorb less incident NIR light than visible (RGB) light, thereby making it difficult to measure the blood volume changes caused by the cardiac cycle by calculating the skin color changes.

Kado *et al.* [30] used a single-plate camera that enabled the simultaneous recording of RGB and NIR information for HR estimation. However, the method [30] had a limitation. When the background lights illuminating the scenes were steady, the use of NIR signals decreased the accuracy of HR estimation, as mentioned above. Conversely, when the background illumination fluctuated significantly, the use of RGB signals decreased the accuracy of the HR estimation.

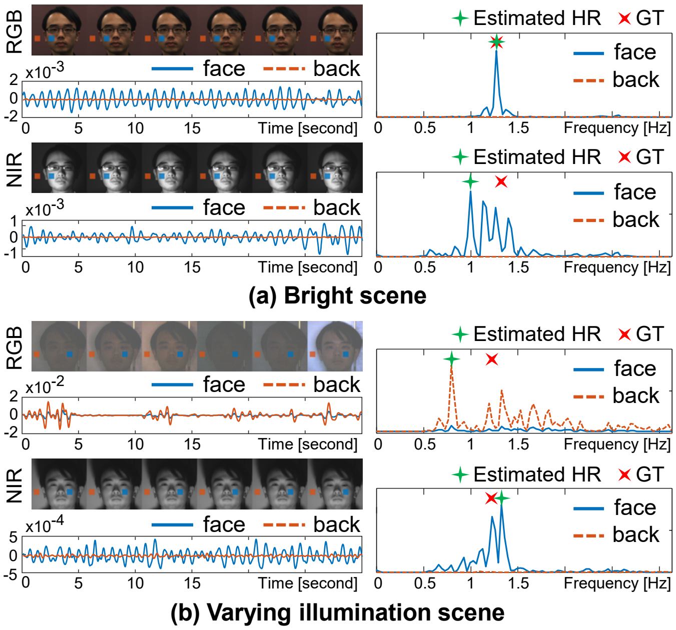
An example of the HR estimates obtained using RGB and NIR videos captured under bright and varying illumination conditions is shown in Fig. 1. For HR estimation using the RGB videos, the green components are the most reliable of the RGB signals [9]; thus, we analyzed the changes in the green pixel values. We measured the temporal changes in the green and NIR pixel values in the face region (blue rectangular region shown in Fig. 1). To investigate the impact of the background illumination on the accuracy of the HR estimation, we also measured the temporal changes in the green and NIR pixel values in the background region (orange rectangular region shown in Fig. 1). We then computed their power spectra using these extracted time-series signals. In line with previous studies [17], [30], we estimated the HR by exploring the frequency with the largest power spectrum component in the power spectra.

In a bright stable scene (Fig. 1 (a)), HR estimation using RGB videos was observed to be more accurate, compared with NIR videos. This may be attributable to the fact that the HR-related RGB face patch signal can be accurately

Kosuke Kurihara and Takayuki Hamamoto are with the Department of Electrical Engineering, Tokyo University of Science, Tokyo 125-8585, Japan.

Daisuke Sugimura is with the Department of Computer Science, Tsuda University, Tokyo 187-8577, Japan (e-mail: sugimura@tsuda.ac.jp).

This article has been accepted for publication in IEEE Transactions on Image Processing (TIP). This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3094739



(a) Bright scene

(b) Varying illumination scene

Fig. 1. Challenges for HR estimation difficulties under various illumination conditions. The time-series signals extracted from the face and background patches (blue and orange rectangles) in each video are shown below each video sequence. The corresponding power spectra are also shown next to each time-series signal. (a) In a bright and stable scene (illumination level: 600 lx), the result obtained using an RGB video was superior to that obtained using an NIR video (ground-truth (GT): 1.26 Hz, RGB: 1.26 Hz, and NIR: 1.00 Hz). (b) In a varying illumination scene (the subject watched an action movie; illumination level: 1 lx), the NIR video yielded a better result than the RGB video (GT: 1.23 Hz, RGB: 0.80 Hz, and NIR: 1.33 Hz).

differentiated from the signal obtained from the background illumination. As shown below in the RGB video sequence, the characteristics of the face- and background-patch time-series signals (blue and orange lines) vary. Conversely, in scenes with significantly varying background illumination (Fig. 1 (b)), the RGB face patch signal was highly influenced by the background illumination variations, as shown in the RGB video sequence below. This indicates that the extraction of the HR-related time-series signal from the RGB video is not accurate. In contrast, we observed that the face patch signal extracted from the NIR video can be observed distinctly from the illumination variations, thereby providing an accurate HR estimate. These preliminary experiments imply that adaptively utilizing RGB and NIR signals can contribute significantly to HR estimation under various illumination conditions.

In this study, we propose a novel method for remote HR estimation using RGB and NIR videos. The key novelty of the proposed method is the incorporation of measures based on cross-correlation analysis between the face and background signals into a framework for HR estimation. As we demonstrated previously, when the background illumination varies significantly, it is preferable to pay greater attention to the NIR signals for accurate HR estimation than to the RGB signals because reflecting NIR light can be captured independently of background illumination variations. Conversely, RGB signals are more reliable than NIR signals for HR estimation when the illumination condition is stable. These observations suggest that correlations between the face and background signals are effective cues for determining a more effective signal between the RGB and NIR for HR estimation. By exploiting the outcomes of the cross-correlation analysis between the face and background signals, the proposed method yields an accurate HR estimation that is robust to illumination variations.

This study is an extended and more detailed version of our earlier study presented in [34]. The new contributions can be summarized as follows. We introduce a scheme that suppresses the influence of extensive subject motions, which significantly degrades the HR estimation accuracy [35], [36]. We report additional experimental results using the following public datasets: TokyoTech Remote PPG [37], MERL-Rice NIR Pulse (MR-NIRP) [12], and UBFC-rPPG [38].

The remainder of this paper is organized as follows: Related works on video-based HR estimation are presented in Section II. In Section III, we provide an overview of the proposed method. In Section IV, we explain the preprocessing procedure for extracting time-series signals related to latent HR. In Section V, we provide a scheme for adaptively fusing RGB and NIR observations, which is the primary contribution of this study. In Section VI, we present the details of motion-robust time-series filtering. In Section VII, we report the results of experiments on real RGB and NIR videos. Finally, we conclude this paper in Section VIII.

II. Related Works

A. RGB Video-based Methods

Poh *et al.* [13] modeled HR signal extraction as a blind signal separation problem. Lam *et al.* [17] proposed a framework for HR estimation based on the majority voting rule for the HR candidates extracted from multiple patch signals. Tulyakov *et al.* [18] introduced self-adaptive matrix completion to eliminate unreliable HR candidates.

Optical and physiological models of light reflection on the skin have been investigated for HR estimation [19], [20], [39]. Haan *et al.* [19] introduced novel chrominance features based on the analysis of a skin reflection model to eliminate specular reflection components that were unrelated to the HR. Based on this model, researchers have proposed an HR estimation method that considers the spectral sensitivity of the camera and the irradiance spectrum of ambient lights [39].

In recent decades, several deep-learning-based methods for HR estimation have been proposed [23], [24], [25], [26], [40], [41], [42]. Spetlík *et al.* [23] proposed a two-step convolutional neural network (CNN) for HR estimation. Chen *et al.* [24] proposed an attention mechanism for the HR estimation. Niu *et al.* [41], [42] constructed spatiotemporal maps using multiple patch signals. Subsequently, for HR estimation, they fed the constructed spatiotemporal maps into a regressor comprising a CNN and a recurrent neural network.

Several researchers have exploited background information for HR estimation to reduce the interference due to background illumination variations [15], [43], [44], [45], [46]. Li *et al.* [15] constructed an adaptive filter using the incoming light reflected from background regions as guidance signals (similar to guided filter [47]). They then applied the filter to the time-series RGB signals observed in the face regions, thereby

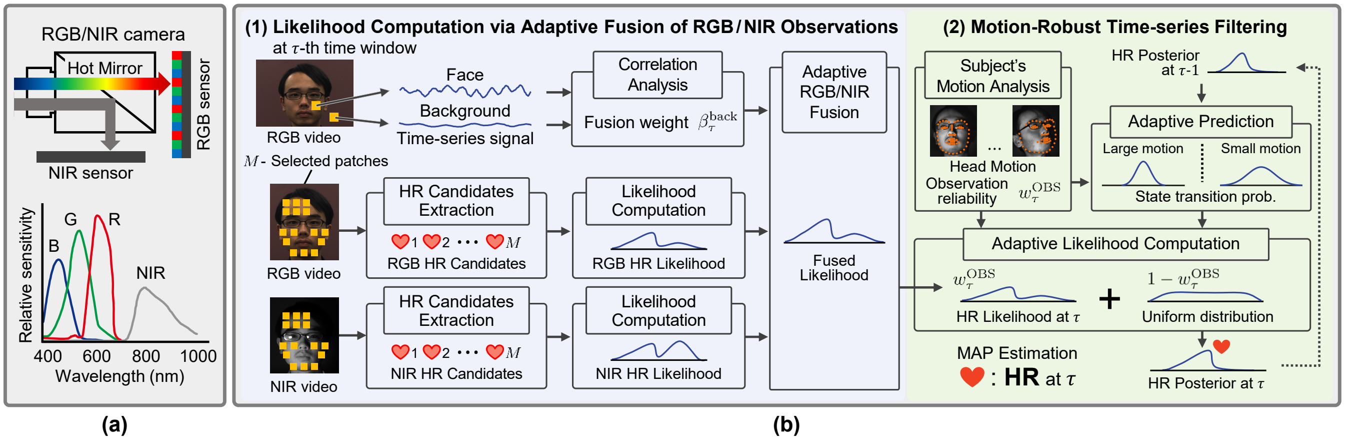


Fig. 2. Overview of our method: (a) Our RGB/NIR imaging system. We utilize two-plate RGB/NIR camera (upper panel). The lower figure shows the relative spectral sensitivity of our imaging system. (b) Our Bayesian HR inference: (1) likelihood computation via adaptive fusion of RGB and NIR observations and (2) motion-robust time-series filtering based on the head movement of the subject.

leading to compensation of the color cast due to background illumination. Tarassenko *et al.* [43] constructed autoregressive models from face and background time-series signals to investigate the frequency correlations between the face and background signals. They then eliminated common frequency components, such as flicker components of fluorescent lights, by using the pole cancelation technique in the z-domain. However, in scenes with largely varying illuminations, latent HR signals are expected to be significantly covered with background illumination, thereby making it difficult to effectively eliminate the components of background illumination variations from the face.

In contrast to the RGB video-based methods so far, we simultaneously exploit both RGB and NIR signals. Because the NIR signals can be captured independently of the visible (RGB) light, the method guarantees the achievement of accurate HR estimation, even under various illumination conditions.

B. NIR Video-based Methods

Gastel *et al.* [28] captured multispectral NIR information to perform accurate HR estimation based on the spectral absorption characteristics of blood in the NIR wavelength band. Park *et al.* [29] deployed a Kalman filter to suppress the impact of the motion of the subjects. Nowara *et al.* [12] exploited a specific NIR band (940 nm), wherein the minimal spectral energy of sunlight emission lights can be observed to perform HR estimation in outdoor environments.

However, HR estimation using NIR videos has been considered more challenging than that using RGB videos because of the physiological characteristics of blood. Therefore, it is more difficult to accurately estimate HR in controlled scenes using NIR video-based methods compared to RGB videos.

C. RGB/NIR Video-based Methods

Kado *et al.* [30] proposed an HR estimation method that simultaneously utilizes RGB and NIR videos. Inspired by the existing RGB video-based method [17], they applied the majority voting rule to HR candidates that were randomly extracted from the RGB and NIR facial videos. However, they did not consider the illumination conditions in their study. As described earlier, HR estimation based on RGB signals is less accurate for uncontrolled scenes. Meanwhile, NIR signals make it difficult to accurately estimate HR in controlled scenes because of the physiological characteristics of blood. Thus, it may be inferred that the method [30] is less applicable in conditions with varying illumination.

Contrary to [30], we adaptively estimate the reliability of the RGB and NIR signals by measuring the correlations between the face and background signals; thus, the proposed method can achieve more accurate HR estimation under various illumination conditions.

III. Proposed Framework

A. Overview

An overview of our method is illustrated in Fig. 2.

1) Our Imaging System: We capture RGB and NIR videos that are spatially and temporally aligned using a two-plate RGB/NIR camera (JAI AD-130GE) and an NIR flash. The relative spectral sensitivity characteristics are presented in Fig. 2(a). The peak sensitivities for the blue, green, red, and NIR sensors appear at 460, 541, 608, and 797 nm, respectively. In addition, the full width at half maximum (FWHM) of the spectral sensitivity for the blue, green, red, and NIR sensors are 90, 93, 92, and 151 nm, respectively. Our RGB/NIR camera records 8-bit RGB raw and NIR images with 1296 × 966 resolution at 30 fps. To obtain an RGB image with full resolution, we applied a demosaicing method [48] to each RGB raw frame.

2) Bayesian Inference: Using the captured RGB/NIR image sequences, we perform HR estimation in a Bayesian inference manner. We compute the likelihood using both RGB and NIR observations (Fig. 2 (b)-(1)), which we adaptively fuse via the cross-correlation analysis of the face and background signals. We then perform time-series filtering of the latent HRs by incorporating the influences of the head movements of the subject (Fig. 2 (b)-(2)). In particular, we utilize a particle filter framework [49]. We finally infer the latent HR based on the maximum a posteriori (MAP) framework.

B. Problem Formulation

Let the latent HR at the τ-th time window be h_τ and a pair of RGB and NIR subsequences observed at the τ-th time window be $\mathbf{z}_\tau = \left(\mathbf{I}_\tau^{\mathrm{RGB}}, \mathbf{I}_\tau^{\mathrm{NIR}}\right)$, where $\mathbf{I}_\tau^{\mathrm{RGB}}$ and $\mathbf{I}_\tau^{\mathrm{NIR}}$ denote the subsequences of the RGB and NIR videos, respectively. The size of each time window is denoted by N. We define all the observations up to the τ-th time window as $\mathcal{Z}_{1:\tau} = (\mathbf{z}_1, \ldots, \mathbf{z}_\tau)$.

The posterior probability of h_τ, defined as $p(h_\tau \mid \mathcal{Z}_{1:\tau})$, can be derived using Bayes rule as

$$\begin{aligned} & p(h_\tau \mid \mathcal{Z}_{1:\tau}) \propto p(\mathbf{z}_\tau \mid h_\tau)\ p(h_\tau \mid \mathcal{Z}_{1:\tau-1}) \\ & = p(\mathbf{z}_\tau \mid h_\tau) \int p(h_\tau \mid h_{\tau-1})\ p(h_{\tau-1} \mid \mathcal{Z}_{1:\tau-1})\ \mathrm{d}h_{\tau-1}\ , \end{aligned} \tag{1}$$

where $p(\mathbf{z}_\tau \mid h_\tau)$, $p(h_\tau \mid h_{\tau-1})$, and $p(h_{\tau-1} \mid \mathcal{Z}_{1:\tau-1})$ denote the likelihood at τ, the state transition probability from $\tau - 1$ to τ, and the posterior probability at $\tau - 1$, respectively.

To ensure the robustness of our HR estimation against varying ambient illumination and motion artifacts due to the head movements of the subject, we model the likelihood $p(\mathbf{z}_\tau \mid h_\tau)$ as

$$p(\mathbf{z}_\tau \mid h_\tau) = w_\tau^{\mathrm{obs}}\, p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)\ +\ (1 - w_\tau^{\mathrm{obs}})\, p_{\mathrm{uni}}(\mathbf{z}_\tau \mid h_\tau)\ , \tag{2}$$

where the weight w_τ^{obs} is computed based on the amount of head movement of the subject. In addition, $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$ and $p_{\mathrm{uni}}(\mathbf{z}_\tau \mid h_\tau)$ denote a likelihood evaluated using RGB and NIR observations and a uniform distribution, respectively. A uniform distribution is adopted to alleviate the influence of unreliable observations arising from head movements of the subject during HR estimation. In such cases, we conjecture that past estimates are more reliable than current estimates. By including a uniform distribution in the likelihood, the effects of the past posterior probability $p(h_{\tau-1} \mid \mathcal{Z}_{1:\tau-1})$ can be directly incorporated into the current estimation. Therefore, this adaptive likelihood model makes our method robust against cases with unreliable observations. The methods for computing $p(\mathbf{z}_\tau \mid h_\tau)$ are detailed in subsequent sections.

According to previous studies [5], [50], HR variations are expected to be small within a short duration. This implies that the state transition probability $p(h_\tau \mid h_{\tau-1})$ can be modeled using a first-order autoregressive model. Hence, we consider that h_τ is drawn from a unimodal Gaussian with mean μ $(= h_{\tau-1})$ and standard deviation σ_τ, as follows:

$$p(h_\tau \mid h_{\tau-1}) = \frac{1}{\sqrt{2\pi\sigma_\tau^2}} \exp\left\{ -\frac{(h_\tau - h_{\tau-1})^2}{2\sigma_\tau^2} \right\}. \tag{3}$$

Based on the MAP estimation framework, we infer the latent HR h_τ^* as follows:

$$h_\tau^* = \underset{h_\tau}{\arg\max}\ \ p(h_\tau \mid \mathcal{Z}_{1:\tau})\ . \tag{4}$$

We summarize the overall process of our method in Algorithm 1. The details of each subalgorithm used in Algorithm 1 are described in the subsequent sections. We also list the typical symbols pertaining to our method in Table I.

Algorithm 1 Proposed HR estimation method

```
Input: RGB and NIR videos $(\mathbf{I}_1^D, \mathbf{I}_2^D, \ldots, \mathbf{I}_T^D)$ $(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ ($T$: total number of time windows)
Initialize: Extract facial image patches $\{\mathbf{p}_{1,j}^{R_2}\}_{j=1}^M$ using Algorithm 2
for $\tau = 1, \ldots, T$ do
    Compute a set of temporal patch sequences $\{\mathbf{p}_{\tau,j}^{R_2}\}_{j=1}^M$ using Algorithm 3
    Extract RGB and NIR HR candidates $\{c_{\tau,j}^{\mathrm{RGB}}\}_{j=1}^M$ and $\{c_{\tau,j}^{\mathrm{NIR}}\}_{j=1}^M$ using Algorithm 4
    Compute $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$ using Algorithm 5
    Compute $p(h_\tau \mid \mathcal{Z}_{1:\tau})$ using Algorithm 6
    Infer $h_\tau^*$ (Eq. (4))
end for
Output: Estimated HR sequence $h_1^*, \ldots, h_T^*$
```

TABLE I
LIST OF TYPICAL SYMBOLS UTILIZED IN OUR FRAMEWORK.

Notation	Definition
h_τ	Latent HR at τ-th time window
$\mathbf{z}_\tau$	Pair of RGB and NIR sequences ($\mathbf{I}_\tau^{\mathrm{RGB}}$ and $\mathbf{I}_\tau^{\mathrm{NIR}}$)
$p(h_\tau \mid h_{\tau-1})$	State transition probability (Eq. (3))
$p(\mathbf{z}_\tau \mid h_\tau)$	Overall likelihood (Eq. (2))
$p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$	Likelihood of $\mathbf{z}_\tau$ (Eq. (9))
$p_{\mathrm{uni}}(\mathbf{z}_\tau \mid h_\tau)$	Uniform distribution
$p_D(\mathbf{I}_\tau^D \mid h_\tau)$	Likelihood of $\mathbf{I}_\tau^D$ ($D \in \{\mathrm{RGB}, \mathrm{NIR}\}$) (Eq. (17))
w_τ^D	Reliability of $\mathbf{I}_\tau^D$ (Eq. (10))
$\beta_\tau^{\mathrm{back}}$	Face/background correlation (Eq. (21))
γ_τ	Cross-domain reliability (Eq. (22))
w_τ^{obs}	Observation reliability (Eq. (26))
ν_τ	Relative motion amount (Eq. (25))
σ_τ	Standard deviation of $p(h_\tau \mid h_{\tau-1})$ (Eq. (27))
n	Face-patch interval used for preprocessing
κ	Hyperparameter used in Bayesian inference

IV. Preprocessing

In this section, we describe the preprocessing procedure for facial videos, which is used to extract the HR-related time-series signals from a face. In previous studies based on both machine learning and non-machine learning-based approaches (e.g., [17], [18], [42], [51]), a similar preprocessing method was adopted for HR estimation.

A. Face Patch Selection

We first extract image patches from an NIR video. This is primarily because, compared with RGB patches, NIR patches are less sensitive to background illumination variations. The RGB and NIR videos are temporally and spatially well aligned because we utilize a two-plate RGB/NIR sensor; thus, the positions of the RGB patches to be extracted are the same as those of NIR patches.

Similar to a previous study [51], we first extract the cheek, jaw, and forehead regions from the face of the subject where the temporal skin color variations attributable to the cardiac pulse can be stably extracted. For the cheek and jaw regions in the t-th frame of τ-th time window, we determine each region of interest (ROI) by constructing facial polygons based on 66 facial landmark positions $\{\mathbf{o}_{\tau,t,i}\}_{i=1}^{66}$ that are estimated using the method [52]. Each index of a set of the facial landmarks identifies the position of the corresponding facial landmark. We define the regions of the left and right cheeks and jaws as polygons, using the subsets of the following facial landmarks as vertices: $\mathbf{r}^{\mathrm{LC}} = \{1, 3, 5, 49, 30\}$ (left cheek), $\mathbf{r}^{\mathrm{RC}} = \{13, 15, 17, 55, 30\}$ (right cheek), $\mathbf{r}^{\mathrm{LJ}} = \{5, 7, 9, 58, 49\}$ (left jaw), $\mathbf{r}^{\mathrm{RJ}} = \{13, 11, 9, 58, 55\}$ (right jaw). We denote each constructed region (polygon) by $\mathbf{R}_{\tau,t}^{R_1}$ ($R_1 \in \{\mathrm{LC},\ \mathrm{RC},\ \mathrm{LJ},\ \mathrm{RJ}\}$). We exclude non-skin regions (e.g., glass, hair, and hat) from $\mathbf{R}_{\tau,t}^{R_1}$ estimated using the face parsing method [53]. Using the estimated pixel-wise skin labels $\mathbf{L}_{\tau,t}^{\mathrm{skin}}$, we extract the skin regions from a video as

$$\tilde{\mathbf{R}}_{\tau,t}^{R_1} = \mathbf{R}_{\tau,t}^{R_1} \odot \mathbf{L}_{\tau,t}^{\mathrm{skin}}, \qquad (5)$$

where $\odot$ denotes a Hadamard product operator.

We determine the forehead regions using $\mathbf{L}_{\tau,t}^{\mathrm{skin}}$. We determine the polygons corresponding to the forehead regions as those above the eyebrows in the skin regions. The facial landmarks corresponding to the eyebrows can be obtained using the method [52] (indices for eyebrows: 18–27). We then divide the estimated forehead region into left and right subregions based on the nose line obtained using the corresponding landmarks (points 18 and 34). We denote the estimated left and right forehead regions as $\tilde{\mathbf{R}}_{\tau,t}^{\mathrm{LF}}$ and $\tilde{\mathbf{R}}_{\tau,t}^{\mathrm{RF}}$, respectively. We add $\tilde{\mathbf{R}}_{\tau,t}^{\mathrm{LF}}$ and $\tilde{\mathbf{R}}_{\tau,t}^{\mathrm{RF}}$ to a set of $\tilde{\mathbf{R}}_{\tau,t}^{R_1}$. The final subregions extracted from the face of the subject are denoted by $\tilde{\mathbf{R}}_{\tau,t}^{R_2}$ ($R_2 \in \{\mathrm{LC},\ \mathrm{RC},\ \mathrm{LJ},\ \mathrm{RJ},\ \mathrm{LF},\ \mathrm{RF}\}$).

Local image patches are uniformly selected from each region $\tilde{\mathbf{R}}_{\tau,t}^{R_2}$ at the n-pixel interval. We define the locations of the selected image patches in the t-th frame of the τ-th time window as $\{\mathbf{p}_{\tau,t,j}^{R_2}\}_{j=1}^{M}$, where M denotes the total number of local patches selected.

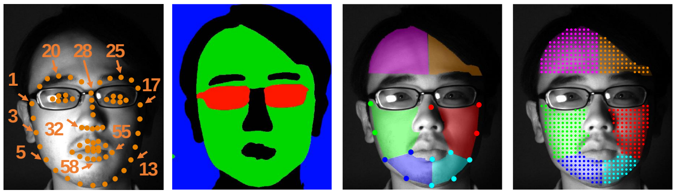
This process is summarized in Algorithm 2. We also show an example of the extracted facial landmarks, face parsing result, each determined subregion in the face region, and selected local patches in Fig. 3 (a), (b), (c), and (d), respectively.

B. Face Patch Tracking

We track each local image patch to obtain the time-series signals corresponding to the skin color changes due to the cardiac pulse in successive frames. Similar to a previous study [51], we assume that the left and right cheeks, jaws, and forehead regions could be approximately represented as planar regions. Therefore, the correspondences of each subregion in consecutive frames can be estimated using projective transformations. To estimate the operators of the projective transformations for these planar regions, we utilize facial landmarks belonging to $\mathbf{r}^{\mathrm{LC}}$, $\mathbf{r}^{\mathrm{RC}}$, $\mathbf{r}^{\mathrm{LJ}}$, and $\mathbf{r}^{\mathrm{RJ}}$. For each left and right forehead region, it is difficult to estimate the projective transformation operators because facial landmarks above the eyebrow cannot be detected. In this method, we apply the operators of projective transformations for the left and right cheek regions to the left and right forehead regions, respectively. This alternative use of projection operators may cause deterioration in tracking accuracy. However, we expect that such deterioration might be insignificant because the anatomical structure of the human face restricts the global movement of the subregions of the face.

Algorithm 2 Face patch selection

```
Input: NIR video frame in $t$-th frame of the $\tau$-th time window
Extract $\{\mathbf{o}_{\tau,t,i}\}_{i=1}^{66}$ using [52]
Estimate $\mathbf{L}_{\tau,t}^{\mathrm{skin}}$ using [53]
Construct $\mathbf{R}_{\tau,t}^{R_1}$ using $\{\mathbf{o}_{\tau,t,i}\}$
Extract $\tilde{\mathbf{R}}_{\tau,t}^{R_1}$ (Eq. (5))
Construct facial polygons of the forehead region
Select $\{\mathbf{p}_{\tau,t,j}^{R_2}\}_{j=1}^{M}$ uniformly from $\tilde{\mathbf{R}}_{\tau,t}^{R_2}$
Output: $M$ image patches $\{\mathbf{p}_{\tau,t,j}^{R_2}\}_{j=1}^{M}$
```



(a) Facial landmarks (b) Semantic labels (c) Region of interest (d) Local patches

Fig. 3. Example of local face patch selection: (a) Facial landmarks extracted using the method [52]. Each number indicates index for corresponding facial landmark; (b) Semantic labels estimated using face parsing [53] (blue: background, green: skin regions, red: eye glass regions) (c) Extracted ROI (green: left cheek, red: right cheek, blue: left jaw, cyan; right jaw, magenta: left forehead, orange: right forehead regions). Circular points represent facial landmarks used to determine each ROI; and (d) Extracted local image patches.

Using the detected facial landmarks $\{\mathbf{o}_{\tau,t,i}\}_{i=1}^{66}$, we compute the projective transformation matrices $\mathbf{H}_{\tau,t}^{R_2 *}$ ($R_2 \in \{\mathrm{LC},\ \mathrm{RC},\ \mathrm{LJ},\ \mathrm{RJ},\ \mathrm{LF},\ \mathrm{RF}\}$) that align each ROI between the t-th and $t+1$-th frame as

$$\mathbf{H}_{\tau,t}^{R_2 *} = \underset{\mathbf{H}_{\tau,t}^{R_2}}{\arg\min} \sum_{i \in \mathbf{r}^{R_2}} \| \mathbf{H}_{\tau,t}^{R_2} \mathbf{o}_{\tau,t,i} - \mathbf{o}_{\tau,t+1,i} \|_2^2 \ . \qquad (6)$$

However, facial occlusions degrade the accuracy of projective transformation, thereby making time-series signals unreliable. To resolve the occlusion problems, we first determine whether the subregions of the face are occluded. When the width or height of the rectangular region that encloses the R_2-th subregion defined as $x_t^{R_2}$ and $y_t^{R_2}$, respectively, is smaller than the patch interval n, we consider that the subregion $\tilde{\mathbf{R}}_{\tau,t}^{R_2}$ is occluded. This processing is represented as

$$\mathrm{Bin}(\tilde{\mathbf{R}}_{\tau,t}^{R_2}) = \begin{cases} 1 & \text{if } (x_t^{R_2} < n) \text{ or } (y_t^{R_2} < n) \\ 0 & \text{otherwise} \ , \end{cases} \qquad (7)$$

where $\mathrm{Bin}(\cdot)$ denotes a binary classification operator.

Using the occlusion detection results, we only apply the transformation matrix $\mathbf{H}_{\tau,t}^{R_2 *}$ to the local patches corresponding to the non-occluded subregions. The patch location in the $t+1$-th frame is represented as

$$\mathbf{p}_{\tau,t+1,j}^{R_2} = \begin{cases} \mathbf{p}_{\tau,t,j}^{R_2} & \text{if } (\mathrm{Bin}(\tilde{\mathbf{R}}_{\tau,t}^{R_2}) = 1) \\ \mathbf{H}_{\tau,t}^{R_2 *} \ \mathbf{p}_{\tau,t,j}^{R_2} & \text{otherwise} \ . \end{cases} \qquad (8)$$

The occluded regions can be accurately tracked once they return to the non-occluded state. This is primarily because we assume that occlusions would be observed only for a short duration. Furthermore, the anatomical structure of the human face restricts the movement of facial landmarks. Therefore, the patch positions corresponding to the subregions classified

Algorithm 3 Face patch tracking in the τ-th time window

```
Input: A set of locations of image patches at the first frame $\{\mathbf{p}^{R_2}_{\tau,1,j}\}$ and the facial landmark sequence $(\mathbf{o}_{\tau,1,i}, \ldots, \mathbf{o}_{\tau,N,i})$
for $t = 1, 2, \ldots, N$ in a $\tau$-th time window do
    for each region in $R_2$ do
        for $j = 1, \ldots, M$ do
            Compute $\mathbf{H}^{R_2 *}_{\tau,t}$ (Eq. (6))
            if $\mathrm{Bin}(\tilde{\mathbf{R}}^{R_2}_{\tau,t}) = 1$ then // occlusion
                $\mathbf{p}^{R_2}_{\tau,t+1,j} \leftarrow \mathbf{p}^{R_2}_{\tau,t,j}$ (Eq. (8))
            else
                $\mathbf{p}^{R_2}_{\tau,t+1,j} \leftarrow \mathbf{H}^{R_2 *}_{\tau,t}\mathbf{p}^{R_2}_{\tau,t,j}$ (Eq. (8))
            end if
        end for
    end for
end for
Output: A set of the temporal patch sequence $\{\mathbf{P}_{\tau,j}\}_{j=1}^{M}$
```

as occluded will not change significantly, which enables our system to track the subregions.

We sequentially perform the above-mentioned processing in subsequent frames. We denote the j-th temporal sequence of the associated patch location between successive frames in the τ-th time window as $\mathbf{P}_{\tau,j} = (\mathbf{p}_{\tau,1,j} \ , \ldots, \mathbf{p}_{\tau,N,j})$. We summarize this procedure in Algorithm 3.

V. Adaptive Fusion of RGB/NIR HR Candidates

We model the observation likelihood term $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$ in Eq. (2) using RGB and NIR observations as follows:

$$p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau) = \frac{w_\tau^{\mathrm{RGB}}}{w_\tau^{\mathrm{RGB}} + w_\tau^{\mathrm{NIR}}}\ p_{\mathrm{RGB}}(\mathbf{I}_\tau^{\mathrm{RGB}} \mid h_\tau) + \frac{w_\tau^{\mathrm{NIR}}}{w_\tau^{\mathrm{RGB}} + w_\tau^{\mathrm{NIR}}}\ p_{\mathrm{NIR}}(\mathbf{I}_\tau^{\mathrm{NIR}} \mid h_\tau), \quad (9)$$

where $p_{\mathrm{RGB}}(\mathbf{I}_\tau^{\mathrm{RGB}} \mid h_\tau)$ and $p_{\mathrm{NIR}}(\mathbf{I}_\tau^{\mathrm{NIR}} \mid h_\tau)$ denote the likelihoods computed by exploiting RGB and NIR videos, respectively. The weights w_τ^{RGB} and w_τ^{NIR} are respectively defined as

$$\begin{aligned} w_\tau^{\mathrm{RGB}} &= \beta_\tau^{\mathrm{back}}\gamma_\tau, && (10) \\ w_\tau^{\mathrm{NIR}} &= (1 - \beta_\tau^{\mathrm{back}})(1 - \gamma_\tau), && (11) \end{aligned}$$

where $\beta_\tau^{\mathrm{back}}$ is computed through a cross-correlation analysis of the signals observed in the face and background regions. The other weight γ_τ assesses which signals measured from the RGB and NIR videos are more reliable. We describe the process of computing these weights in the subsequent sections.

A. HR Candidates Extraction from RGB Video

To compute $p_{\mathrm{RGB}}(\mathbf{I}_\tau^{\mathrm{RGB}} \mid h_\tau)$, we first extract the RGB signals from each patch in $\mathbf{P}_{\tau,j}$ at every frame. The extracted RGB signals are defined as $\{\mathbf{s}^c_{\tau,j}\}_{c \in \{\mathrm{R,G,B}\}}$, where each component $\mathbf{s}^c_{\tau,j}$ denotes a time-series signal in the $c\,(\in \{\mathrm{R, G, B}\})$ -th color channel: $\mathbf{s}^c_{\tau,j} = (s^c_{\tau,1,j}, \ldots, s^c_{\tau,N,j})$.

1) Color-Difference Space Conversion: Our method processes the RGB signals in a color-difference space for RGB HR candidate extraction. This is primarily because analysis in a color-difference space enables performance enhancement for RGB video-based HR estimation [19], [20]. Notably, it has been reported that the color-difference-based method [20] enables superior performance in HR estimation in scenes where stable and sufficient illumination is provided [54].

We project the RGB values $\{\mathbf{s}^c_{\tau,j}\}_{c \in \{\mathrm{R,G,B}\}}$ onto a color-difference space. Following the method [20], the color-space conversion can be represented as

$$\begin{aligned} \mathbf{t}^{\mathrm{p1}}_{\tau,j} &= \mathbf{s}^{\mathrm{G}}_{\tau,j} - \mathbf{s}^{\mathrm{B}}_{\tau,j}, && (12) \\ \mathbf{t}^{\mathrm{p2}}_{\tau,j} &= -2\mathbf{s}^{\mathrm{R}}_{\tau,j} + \mathbf{s}^{\mathrm{G}}_{\tau,j} + \mathbf{s}^{\mathrm{B}}_{\tau,j}. && (13) \end{aligned}$$

Using the projected components $\mathbf{t}^{\mathrm{p1}}_{\tau,j}$ and $\mathbf{t}^{\mathrm{p2}}_{\tau,j}$, the chrominance components $\mathbf{d}_{\tau,j}$ are obtained as

$$\mathbf{d}_{\tau,j} = \mathbf{t}^{\mathrm{p1}}_{\tau,j} + \frac{\mathrm{Std}(\mathbf{t}^{\mathrm{p1}}_{\tau,j})}{\mathrm{Std}(\mathbf{t}^{\mathrm{p2}}_{\tau,j})}\ \mathbf{t}^{\mathrm{p2}}_{\tau,j}, \quad (14)$$

where $\mathrm{Std}(\mathbf{t})$ denotes an operator that computes the standard deviation of a time-series signal $\mathbf{t}$.

2) Frequency Analysis: We analyze the spectral characteristics of the extracted chrominance components $\mathbf{d}_{\tau,j}$. Similar to previous studies [17], [30], [35], we apply a band-pass filter to $\mathbf{d}_{\tau,j}$ to remove the frequency components outside the range of the HR. We determine the bandwidth of the band-pass filter to be 0.7 – 4 Hz on the basis of prior knowledge, according to the normal HR range [55]. Subsequently, we perform a fast Fourier transform (FFT) to obtain the power spectrum $\mathbf{u}^{\mathrm{RGB}}_{\tau,j}$ of the filtered signals. This processing is represented as

$$\mathbf{u}^{\mathrm{RGB}}_{\tau,j} = \Phi(\,\mathrm{BPF}(\mathbf{d}_{\tau,j})\,), \quad (15)$$

where $\mathrm{BPF}(\cdot)$ denotes an operator of the band-pass filter and $\Phi(\cdot)$ denotes a function that computes the power spectrum of an input signal via the FFT.

3) HR Candidates Extraction: We explore the largest power spectrum component v_1 in $\mathbf{u}^{\mathrm{RGB}}_{\tau,j}$. We denote the frequency with v_1 as $c^{\mathrm{RGB}}_{\tau,j}$, which can be assumed to be related to the cardiac pulse of a person [13], [17]. Following [17], [30], we introduce a confidence score that assesses the reliability of the HR candidate $c^{\mathrm{RGB}}_{\tau,j}$. In addition, we consider that local patches belonging to the subregions classified as occlusions (Eq. (7)) are unreliable; thus, we set the confidence scores corresponding to these local patches to zero. Therefore, the confidence score $\alpha^{\mathrm{RGB}}_{\tau,j}$ is computed as

$$\alpha^{\mathrm{RGB}}_{\tau,j} = \begin{cases} 0 & \text{if } (j \in \tilde{\mathbf{R}}^{R_2}_{\tau,t}\ , \sum_{t \in \tau} \mathrm{Bin}(\tilde{\mathbf{R}}^{R_2}_{\tau,t}) \geq 1) \\ v_1/v_2 - 1 & \text{otherwise}\ , \end{cases} \quad (16)$$

where v_2 denotes the second-largest power spectrum component in $\mathbf{u}^{\mathrm{RGB}}_{\tau,j}$.

Finally, we obtain a pair of HR candidates and the corresponding confidence scores for all the patches, which are denoted by $(c^{\mathrm{RGB}}_{\tau,1}, \ldots, c^{\mathrm{RGB}}_{\tau,M})$ and $(\alpha^{\mathrm{RGB}}_{\tau,1}, \ldots, \alpha^{\mathrm{RGB}}_{\tau,M})$, respectively. This process is summarized in Algorithm 4.

Algorithm 4 HR candidates extraction from RGB video

```
Input: A set of the temporal patch sequences $\{\mathbf{P}_{\tau,j}\}_{j=1}^{M}$ and
    the RGB video $\mathbf{I}_{\tau}^{\mathrm{RGB}}$
for $j = 1, \ldots, M$ do
    Extract $\mathbf{s}_{\tau,j}^{c}$ from $\mathbf{P}_{\tau,j}$ in $\mathbf{I}_{\tau}^{\mathrm{RGB}}$
    Extract $\mathbf{d}_{\tau,j}$ from $\mathbf{s}_{\tau,j}^{c}$ (Eq. (14))
    Compute $\mathbf{u}_{\tau,j}^{\mathrm{RGB}}$ (Eq. (15))
    Estimate $c_{\tau,j}^{\mathrm{RGB}}$ and $\alpha_{\tau,j}^{\mathrm{RGB}}$ (Eq. (16))
end for
Output: $M$ pairs of the HR candidates and the corresponding
    confidence scores $\{c_{\tau,j}^{\mathrm{RGB}}, \alpha_{\tau,j}^{\mathrm{RGB}}\}_{j=1}^{M}$
```

B. HR Candidates Extraction from NIR Video

RGB-based methods are prone to produce inaccurate HR estimates in low-light scenes [54] and in scenes where the spectral composition of the light source (i.e., ambient light) is variable [20]. In such unstable illumination scenes, NIR processing enables the extraction of signals attributable to the HR more accurately compared with the RGB. Here, we detail the extraction of HR candidates from an NIR video.

First, we extract the NIR signals $\mathbf{s}_{\tau,j}^{\mathrm{NIR}} = (s_{\tau,1,j}^{\mathrm{NIR}}, \ldots, s_{\tau,N,j}^{\mathrm{NIR}})$ from each patch sequence in the NIR video. Then, we apply frequency analysis (Section V-A2) to each local NIR patch signal to obtain its power spectrum $\mathbf{u}_{\tau,j}^{\mathrm{NIR}}$. We obtain a pair of HR candidates and the corresponding confidence scores for all the selected patches and denote them as $(c_{\tau,1}^{\mathrm{NIR}}, \ldots, c_{\tau,M}^{\mathrm{NIR}})$ and $(\alpha_{\tau,1}^{\mathrm{NIR}}, \ldots, \alpha_{\tau,M}^{\mathrm{NIR}})$, respectively.

The NIR processing described so far enables the extraction of HR candidates with minimal spectrum noise. Because our imaging system employs an NIR flash unit to sufficiently illuminate the face region of the subject, the reflecting NIR light can be captured with less noise. In addition, it can be assumed that the NIR amount of ambient light will be small in indoor scenes. This assumption has been widely accepted in the field of image processing, for tasks such as image enhancement using a pair of RGB and NIR images [56], [57], [58], [59], [60], [61], [62].

C. Likelihood Computation

We model the RGB and NIR likelihoods $p_{\mathrm{RGB}}(\mathbf{I}_{\tau}^{\mathrm{RGB}} \mid h_{\tau})$ and $p_{\mathrm{NIR}}(\mathbf{I}_{\tau}^{\mathrm{NIR}} \mid h_{\tau})$ using the weighted kernel density estimation. We compute $p_D(\mathbf{I}_{\tau}^{D} \mid h_{\tau})$ $(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ as

$$p_D(\mathbf{I}_{\tau}^{D} \mid h_{\tau}) = \frac{1}{\sum_{j=1}^{M} \alpha_{\tau,j}^{D}} \frac{1}{MW} \sum_{j=1}^{M} \alpha_{\tau,j}^{D} K\left(\frac{h_{\tau} - c_{\tau,j}^{D}}{W}\right) , \quad (17)$$

where $K(\cdot)$ denotes a Gaussian kernel with bandwidth W. Considering the sampling rate of a video f_s and the size of the time window N, we set W as $f_s/2N$.

D. Analyzing Face/Background Correlations

We perform a cross-correlation analysis of the face and background regions. When the background lights provide stable and sufficient illumination, we expect that the signals relevant to the HR will be distinguishable from the background illuminations, thereby indicating weak correlations between the face and background signals. Conversely, when the background light fluctuates significantly, a dominant color cast occurs on the face regions owing to the varying background illumination, indicating strong correlations between the face and background signals. Based on these assumptions, we compute weight $\beta_{\tau}^{\mathrm{back}}$ using the results of the cross-correlation analysis of the face and background signals.

1) Averaged Face/Background Signal Extraction: We utilize the green channel of the RGB image to analyze correlations between the face and background regions because of the visible light absorption characteristics of blood. According to previous studies [32], [33], (oxy-) hemoglobin absorbs green light more than red light and penetrates deeper into the skin compared with blue light to probe the vasculature. The authors in [33] showed the results for the spectral blood volume pulse signal-to-noise ratios (SNRs). Referring to the discussions in Section 5 and Fig. 4 in [33], the peak SNR response appears at 578 nm, and a region with a relative SNR within -10 dB of the peak occurs from 512 to 609 nm. This indicates that the green component will include signals relevant to the HR more than the red and blue components. Conversely, the green lights reflected from the background regions do not contain signals attributable to the HR. Hence, the green component of the reflecting light is a reliable indicator for analyzing the correlation between the face and background regions.

We calculate the face signal $\mathbf{f}_{\tau}$ using the green signals of the local patches $\{\mathbf{s}_{\tau,j}^{\mathrm{G}}\}_{j=1}^{M}$ as follows:

$$\mathbf{f}_{\tau} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{s}_{\tau,j}^{\mathrm{G}} \ . \quad (18)$$

We also define a green signal in the background regions $\mathbf{b}_{\tau}$ that is previously estimated using face parsing [53] (Section IV-A). We uniformly select M number of local image patches from the background regions. Similar to computing $\mathbf{f}_{\tau}$, we compute $\mathbf{b}_{\tau}$ by averaging the green signals observed in the local patches in the background regions as

$$\mathbf{b}_{\tau} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{v}_{\tau,j}^{\mathrm{G}} \ , \quad (19)$$

where $\mathbf{v}_{\tau,j}^{\mathrm{G}}$ denotes the green signal observed in the j-th local patch in the background region.

2) Face/Background Correlation Analysis: To compute the weight $\beta_{\tau}^{\mathrm{back}}$, we measure the correlation between $\mathbf{f}_{\tau}$ and $\mathbf{b}_{\tau}$. We first apply the band-pass filter (0.7 – 4 Hz) to $\mathbf{f}_{\tau}$ and $\mathbf{b}_{\tau}$. We denote the filtered face and background signals as $\mathrm{BPF}(\mathbf{f}_{\tau})$ and $\mathrm{BPF}(\mathbf{b}_{\tau})$, respectively. We then compute the Pearson correlation ρ_{τ} between $\mathrm{BPF}(\mathbf{f}_{\tau})$ and $\mathrm{BPF}(\mathbf{b}_{\tau})$ as

$$\rho_{\tau} = \frac{\mathrm{Cov}(\,\mathrm{BPF}(\mathbf{f}_{\tau}), \mathrm{BPF}(\mathbf{b}_{\tau})\,)}{\mathrm{Std}(\mathrm{BPF}(\mathbf{f}_{\tau}))\,\mathrm{Std}(\mathrm{BPF}(\mathbf{b}_{\tau}))}, \quad (20)$$

where $\mathrm{Cov}(\mathbf{f}, \mathbf{b})$ denotes an operator that computes the covariance between the time-series signals $\mathbf{f}$ and $\mathbf{b}$.

Using the computed ρ_{τ}, we calculate $\beta_{\tau}^{\mathrm{back}}$ as follows:

$$\beta_{\tau}^{\mathrm{back}} = 1 - \rho_{\tau} \ . \quad (21)$$

Algorithm 5 Adaptive fusion of RGB/NIR HR candidates

```
Input: HR candidates and corresponding confidence scores $\{c^D_{\tau,j}, \alpha^D_{\tau,j}\}^M_{j=1}$, RGB and NIR videos $\mathbf{I}^D_\tau$ ($D \in \{\mathrm{RGB}, \mathrm{NIR}\}$), and green patch signals $\{\mathbf{s}^{\mathrm{G}}_{\tau,j}\}^M_{j=1}$.
Compute $p_D(\mathbf{I}^D_\tau \mid h_\tau)$ using $\{c^D_{\tau,j}, \alpha^D_{\tau,j}\}^M_{j=1}$ (Eq. (17))
Compute $\rho_\tau$ (Eq. (20))
Compute $\beta^{\mathrm{back}}_\tau$ using $\rho_\tau$ (Eq. (21))
Compute $\gamma_\tau$ (Eq. (22))
Compute $w^{\mathrm{RGB}}_\tau$ and $w^{\mathrm{NIR}}_\tau$ using $\beta^{\mathrm{back}}_\tau$ and $\gamma_\tau$ (Eq. (10))
Compute $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$ (Eq. (9))
Output: Observation likelihood $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$
```

In varying illumination scenes, ρ_τ has a large value because the face signals are highly correlated with the background illumination. Therefore, we assign a high confidence score ($\beta^{\mathrm{back}}_\tau$) to the NIR likelihood $p_{\mathrm{NIR}}(\mathbf{I}^{\mathrm{NIR}}_\tau \mid h_\tau)$. Conversely, in bright scenes, ρ_τ is assigned a low value (approximately 0) because the time-series signals in the face and background regions are correlated to a lesser degree. Accordingly, we assign a high confidence score ($\beta^{\mathrm{back}}_\tau$) to the RGB likelihood $p_{\mathrm{RGB}}(\mathbf{I}^{\mathrm{RGB}}_\tau \mid h_\tau)$. This makes it possible to fuse the RGB and NIR observations adaptively based on the background illumination conditions.

E. Cross-domain Reliability Computation

We assess which candidate is more reliable for HR estimation, the RGB or NIR? When the confidence scores of the RGB videos $\{\alpha^{\mathrm{RGB}}_{\tau,j}\}$ are higher than those of the NIR videos $\{\alpha^{\mathrm{NIR}}_{\tau,j}\}$, we consider the RGB patch signals to be more reliable for HR estimation. Because the confidence score is calculated using the ratio of the largest and the second-largest power spectrum components, a high confidence score indicates that the power spectrum component related to the HR is distinctly observed. Furthermore, RGB signals are more reliable than NIR signals for HR estimation because of the physiological characteristics of the blood. Therefore, the RGB HR candidates $\{c^{\mathrm{RGB}}_{\tau,j}\}$ with high confidence scores are more ideal for HR estimation than NIR ones. We use this idea to compute the cross-domain reliability γ_τ.

To compute γ_τ, we first compute each mode (i.e., the most frequently observed value) among a set of HR candidates $\{c^{\mathrm{RGB}}_{\tau,j}\}$ and $\{c^{\mathrm{NIR}}_{\tau,j}\}$, defined as $\tilde{c}^{\mathrm{RGB}}_\tau$ and $\tilde{c}^{\mathrm{NIR}}_\tau$, respectively. Using the confidence scores of the patches belonging to the modes $\tilde{c}^{\mathrm{RGB}}_\tau$ and $\tilde{c}^{\mathrm{NIR}}_\tau$, we compute γ_τ as

$$\gamma_\tau = \frac{\sum_{j \in \Omega^{\mathrm{RGB}}_\tau} \alpha^{\mathrm{RGB}}_{\tau,j}}{\sum_{j \in \Omega^{\mathrm{RGB}}_\tau} \alpha^{\mathrm{RGB}}_{\tau,j} + \sum_{j \in \Omega^{\mathrm{NIR}}_\tau} \alpha^{\mathrm{NIR}}_{\tau,j}} \ , \tag{22}$$

where $\Omega^{\mathrm{RGB}}_\tau$ and $\Omega^{\mathrm{NIR}}_\tau$ denote the sets of patch indices belonging to the modes $\tilde{c}^{\mathrm{RGB}}_\tau$ and $\tilde{c}^{\mathrm{NIR}}_\tau$, respectively.

We summarize our scheme for adaptive fusion of RGB/NIR likelihoods in Algorithm 5.

VI. Motion-Robust Time-series Filtering

We perform motion-robust time-series filtering on the HR to address the head movements of subjects, which significantly decrease the HR estimation accuracy. When a subject moves considerably, it is difficult to extract reliable HR-related time-series signals [35]. This is primarily because the significant changes in the reflective light (e.g., shadow and specular) due to head movement destabilize the RGB/NIR observations. To alleviate such problems, we introduce a reliability indicator w^{obs}_τ into our time-series filtering to determine the stability of the current observations. We adaptively control the shape of the likelihood $p(\mathbf{z}_\tau \mid h_\tau)$ and state transition probability $p(h_\tau \mid h_{\tau-1})$ using w^{obs}_τ.

A. Computation of Observation Reliability

When significant head movements are observed, we assume that the past estimates are more reliable than the current estimates, which are computed using unreliable observations. We assume that the current observations will be reliable again upon the cessation of head movements; therefore, the current observations are preferable to be particularly exploited for posterior distribution computation.

We compute the observation reliability as follows. First, we measure the amount of head movements using the trajectories of all the patches. Specifically, we compute the mean subject-head trajectory $\mathbf{Q}_\tau = (\mathbf{q}_{\tau,1}, \ldots, \mathbf{q}_{\tau,N})$ in the τ-th time window by aggregating the trajectories of all facial patches $\{\mathbf{P}_{\tau,j}\}$. The mean position of the subject's head at the t-th frame in the τ-th time window $\mathbf{q}_{\tau,t}$ is represented as

$$\mathbf{q}_{\tau,t} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{p}_{\tau,t,j} \ . \tag{23}$$

Using $\mathbf{q}_{\tau,t}$, we compute the cumulative motion amount in the τ-th window m_τ as

$$m_\tau = \sum_{t=2}^{N} \|\mathbf{q}_{\tau,t} - \mathbf{q}_{\tau,t-1}\|_2 \ . \tag{24}$$

To measure the changes in the current motion amount relative to past observations, we compute the relative motion amount in the τ-th time window ν_τ as

$$\nu_\tau = \frac{m_\tau}{\sum_{\tau'=1}^{\tau-1} m_{\tau'}/(\tau-1) + m_\tau} \ . \tag{25}$$

When the head movement of the subject is significant, compared with the past time windows, ν_τ reaches 1, indicating that the current observations are unreliable. When the head movements are less expansive, ν_τ becomes close to 0.

We compute the reliability of the current observations w^{obs}_τ using ν_τ as follows:

$$w^{\mathrm{obs}}_\tau = 1 - \nu_\tau \ . \tag{26}$$

Fusing $p_{\mathrm{obs}}(\mathbf{z}_\tau \mid h_\tau)$ and $p_{\mathrm{uni}}(\mathbf{z}_\tau \mid h_\tau)$ with w^{obs}_τ (Eq. (2)) enables us to adaptively control the contributions of the current observations to compute the posterior probability.

B. Adaptive Control of State Transition Probability

We adaptively control the shape of the state transition probability $p(h_\tau \mid h_{\tau-1})$ using w^{obs}_τ. When we observe significant head movements, we regulate the state transition probability

Algorithm 6 Motion-robust time-series filtering

```
Input: Likelihood $p_{\mathrm{obs}}(\mathbf{z}_\tau \,|\, h_\tau)$, previous posterior distribution $p(h_{\tau-1} \,|\, \mathcal{Z}_{1:\tau-1})$, a set of the temporal patch sequences $\{\mathbf{P}_{\tau,j}\}_{j=1}^{M}$, and the motion amount sequence $(m_1, \ldots, m_{\tau-1})$
    Compute $m_\tau$ (Eqs. (23) and (24))
    Compute $\nu_\tau$ using $(m_1, \ldots, m_\tau)$ (Eq. (25))
    Compute $w_\tau^{\mathrm{obs}}$ using $\nu_\tau$ (Eq. (26))
    Compute $p(\mathbf{z}_\tau \,|\, h_\tau)$ (Eq. (2))
    Control $p(h_\tau \,|\, h_{\tau-1})$ using $\nu_\tau$ (Eqs. (3) and (27))
    Compute $p(h_\tau \,|\, \mathcal{Z}_{1:\tau})$ (Eq. (1))
Output: Posterior distribution $p(h_\tau \,|\, \mathcal{Z}_{1:\tau})$
```

to be a narrow distribution to preserve the influence of the past posterior probability. Conversely, as the head movement of the subject becomes less expansive, we conclude that the observations return to a reliable state. Therefore, at this time, we make the state transition probability a wide distribution such that it can significantly accommodate the influences of the current observations. Accordingly, we control the value of the standard deviation σ_τ, which is used for the state transition probability (Eq. (3). Specifically, we compute σ_τ as

$$\sigma_\tau = \kappa w_\tau^{\mathrm{obs}} \ , \tag{27}$$

where $\kappa > 0$ denotes a parameter. When large head movements are observed, σ_τ is set close to zero, resulting in a narrow distribution of the state transition probability. Conversely, as the head movement becomes less expansive, σ_τ approaches κ; consequently, the state transition probability becomes a wide distribution.

The motion-robust time-series filtering scheme is outlined in Algorithm 6.

VII. Experiments

To demonstrate the effectiveness of our method, We conducted experiments on the following datasets: our RGB/NIR dataset, TokyoTech Remote PPG dataset [37], MR-NIRP dataset [12], and UBFC-rPPG dataset [38].

A. Dataset

We briefly describe all the datasets used in this experiment. We refer to the dataset [37] as "Tokyo," dataset [12] as "MR," and dataset [38] as "UBFC." The details of these datasets are summarized in Table II. To highlight the differences in the amount of head movements of the subjects between the datasets, we show the histogram of the motion amount m_τ in Fig. 4. In this figure, "S1," ..., "S6" refer to the conditions for the sequences in our dataset. We separately computed the histogram of "S5" and "S6" and that of "S1," "S2," "S3" and "S4" because "S5" and "S6" includes the significant head movements of the subjects compared to the other ones. The details for each condition in our dataset are described in the subsequent section. We also show the spectrogram of the background illumination variation averaged over the sequences in each dataset in Fig. 5 to clearly state the differences in the illumination conditions between the datasets.

TABLE II
Details of datasets used in experiments.

	Our Dataset	Tokyo [37]	MR [12]	UBFC [38]
Camera	Two-plate RGB/NIR	Single-chip RGB/NIR	RGB and NIR	RGB
# Subjects	18	9	8	47
# Videos	66	9	8	50
Resolution	1296×966	640×480	640×640	640×480
Frame rate	30 fps	30 fps	30 fps	30 fps
Duration	120 s	180 s	180 s	60 s
Illumination	Bright, Low, Varying	Bright	Bright	Bright
Large motion	Yes (S5 and S6)	No	No	No

1) Our RGB/NIR Video Dataset: We captured the RGB and NIR videos under the following conditions: (1) bright scene (illumination level: 600 lx), (2) low-light scene (illumination level: 0.4 lx), (3) scene under varying illumination with a frequency near the normal HR (illumination level: 1 lx), (4) realistic scene, such as a theater (low-light scene with varying illuminations; illumination level: 1 lx), (5) bright scene with expansive head movements (illumination level: 600 lx), and (6) a varying illumination scene with expansive head movements (illumination level: 2 lx). We refer to these conditions as "S1," "S2," "S3," "S4," "S5," and "S6," respectively. In "S3," we played a movie, and the green lights emitted by the display were sinusoidally oscillated at 0.83 Hz (50 bpm). In "S4," we played a trailer of an action movie. In "S5" and "S6," the participants were instructed to move freely in the field of view of our RGB/NIR camera. The major difference between "S5" and "S6" is the illumination conditions. In "S5," the illumination was stable and bright. Conversely, illumination was sinusoidally oscillated at 0.83 Hz (50 bpm) under a low-light level in "S6." In the other conditions ("S1" – "S4"), the participants were instructed to sit still near the camera. Altogether, eighteen subjects participated, and the dataset contained 66 video recordings. We captured RGB and NIR videos for 2 min using our RGB/NIR camera (JAI AD-130GE) described in Sect. III-A1. A pulse oximeter (CONTEC CMS50D+) was used to obtain the ground-truth HR.

2) TokyoTech Remote PPG Dataset: This dataset contained the RGB and NIR videos of nine subjects, who were instructed to sit still and perform a handgrip exercise for approximately 1 min. They were also instructed to raise one hand. The RGB raw and NIR videos were captured for 3 min using a single-chip RGB/NIR camera under visible and NIR illumination. Notably, we observed several saturated regions in the RGB videos of four subjects (#: 5, 6, 7, and 9). In such scenes, it is difficult for all existing methods to achieve accurate HR estimation because they do not assume heavy pixel saturations in a video. To clearly observe the differences in the HR estimation performance among the existing methods, we excluded saturated videos from the comparison evaluations.

3) MR-NIRP Dataset: This dataset contained the RGB and NIR videos of eight subjects. The RGB and NIR videos were separately captured for 3 min using an RGB camera and an NIR camera under visible and NIR illumination. Because the RGB and NIR videos in this dataset were not spatially aligned, we independently extracted the RGB and NIR local patches from the RGB and NIR videos and used them for the existing

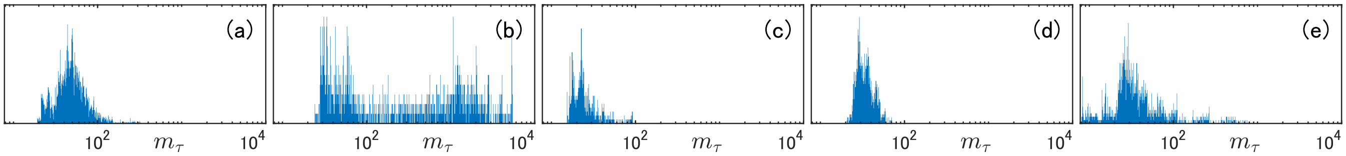


Fig. 4. Normalized histograms of head movements of subjects $m_\mathcal{T}$: (a) "S1," "S2," "S3," and "S4" in our dataset (mean of $m_\mathcal{T}$: 53); (b) "S5" and "S6" in our dataset (mean of $m_\mathcal{T}$: 855); (c) Tokyo [37] (mean of $m_\mathcal{T}$: 27); (d) MR [12] (mean of $m_\mathcal{T}$: 35); (e) UBFC [38] (mean of $m_\mathcal{T}$: 55).

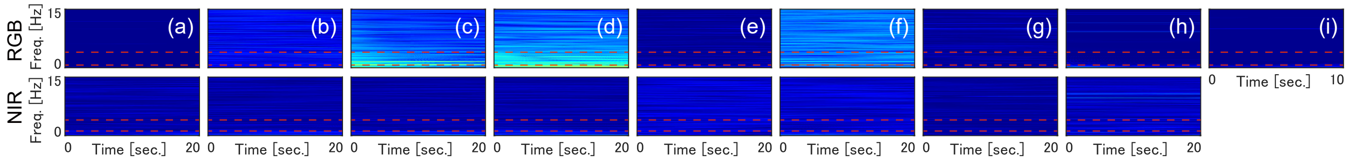


Fig. 5. Averaged spectrograms of RGB and NIR signals extracted from background regions: (a) "S1"; (b) "S2"; (c) "S3"; (d) "S4"; (e) "S5"; (f) "S6"; (g) Tokyo [37]; (h) MR [12]; (i) UBFC [38]. The red dashed line in each spectrogram represents the normal HR range (0.7 - 4 Hz).

methods and our method.

4) UBFC-rPPG Dataset: This dataset contained 50 RGB videos of forty-seven subjects. The subjects of the dataset sat and played a time-sensitive mathematical game under bright and stable illumination conditions. Each video was recorded for 1 min using an RGB camera at 30 fps with 640×480 resolution in an uncompressed 8-bit RGB format. We utilized 49 RGB videos in this dataset for the experiments because one video file could not be read.

B. Evaluation Metrics

We quantitatively evaluated the results using the root mean squared error (RMSE) and mean absolute error (MAE). Similar to previous works [17], [63], [64], we also evaluated the success rate (SR) of the HR estimation by aggregating the outputs for which the difference between the estimated and ground-truth HRs was less than a certain threshold (± 5 bpm). Furthermore, we assessed the HR estimation performance using the Bland-Altman analysis [65], [66], a data-plotting method for evaluating the agreement between the estimated and ground-truth HRs; the plots in which the measurements are narrowly distributed around zero exhibit better performance.

C. Comparison Methods

We compared our method with the following state-of-the-art methods based on non-machine learning: RGB video-based methods ([15], [17], [18], [20]), NIR video-based methods ([12], [27]), and RGB/NIR video-based methods ([30] and our earlier method [34]). According to [20], a band-pass filter that removes the frequency components outside the range of the HR is not utilized. To observe the impact of using the band-pass filter, we incorporated the band-pass filter into [20] and evaluated its performance. We refer to this as "[20] w/ BPF." The method [18] adopted the chrominance feature proposed in the literature [19]. According to [19], two types of chrominance features, represented with and without a band-pass filter, are presented. In this comparison, we investigated the impacts of these differing chrominance features (i.e., the effects of a band-pass filter) on the HR estimation performance of [18]. We refer to them as "[18] w/ BPF" and "[18] w/o BPF," respectively. Note that we implemented these existing methods in accordance with the reference papers. Furthermore, we tested our method using only RGB or NIR signals to observe the impact of using either one for HR estimation. We refer to these methods as "Ours w/o NIR" and "Ours w/o RGB," respectively.

We also compared our method with a machine learning-based method using CNNs [23] on all datasets. In addition, we compared the results obtained by other deep learning-based methods [26], [40] on "UBFC." Note that we listed the results of [26], [40], as reported in the respective papers.

Based on the preliminary experiments, the parameters required for the compared methods were set as follows. Considering the frequency resolution in the frequency analysis, we set $N = 900$ (30 s). We ensured the control parameters of the existing methods were optimal. Our method has two specific parameters to set. The first is the face patch interval n, which is used for preprocessing facial videos, and the other is the κ used for our Bayesian inference. We set these parameters as follows. We set n such that the face patches could be extracted uniformly from a face; specifically, $n = 20$. Based on preliminary experiments, we observed that the choice of n did not affect the quality of the HR estimation performance of our method. For our Bayesian inference, we determined $\kappa = 5$. In practice, setting κ is important because it is used in our Bayesian inference, which is the main part of the proposed method. Hence, we analyzed the dependence of the choice of κ, and the results are presented in Sect. VII-E3.

D. Results

1) Results for Our Dataset: The comparison results for our dataset are presented in Table III. Each value was obtained by averaging the results of all the subjects. It can be observed that our method produced quantitatively better results than the other methods. The results based on the Bland-Altman plot using our dataset are presented in Fig. 6. The plots obtained using our method were narrowly distributed around zero, indicating superior performance over the other methods.

We discuss about the results more in-depth. It was observed that the RGB-based methods showed results comparable with that of our method for "S1" and "S5" at brightly illuminated

TABLE III
QUANTITATIVE COMPARISONS IN OUR DATASET BASED ON THE AVERAGE RMSE, MAE AND SR. THE BEST SCORES ARE PRESENTED IN **BOLD**.

		RMSE (bpm)							MAE (bpm)							SR (%)						
		S1	S2	S3	S4	S5	S6	Avg.	S1	S2	S3	S4	S5	S6	Avg.	S1	S2	S3	S4	S5	S6	Avg.
RGB	[17]	7.27	28.71	22.19	25.08	23.20	25.85	22.05	3.99	21.40	22.08	24.01	17.38	25.58	19.07	83.7	15.6	0.0	4.8	42.5	0.0	24.4
	[15]	3.43	76.18	22.20	20.00	11.78	25.85	26.57	1.63	63.31	22.08	14.40	7.31	25.58	22.38	90.5	7.9	0.0	28.7	64.1	0.0	31.9
	[18] w/o BPF	1.96	133.22	22.20	20.62	10.31	31.01	36.55	0.85	86.22	22.08	19.25	5.41	30.71	27.42	95.3	2.5	0.0	10.5	80.1	0.0	31.4
	[18] w/ BPF	1.30	63.23	22.19	17.73	4.64	36.74	24.30	0.68	50.88	22.07	13.57	2.22	32.08	20.25	95.9	8.9	0.0	26.7	89.9	3.4	37.5
	[20]	1.30	414.73	22.20	25.29	6.62	159.73	104.98	**0.67**	335.50	22.08	21.67	2.91	85.35	78.03	96.1	3.0	0.0	6.6	88.9	17.6	35.4
	[20] w/ BPF	1.31	74.14	22.19	30.48	3.51	62.19	32.30	**0.67**	61.59	22.07	19.52	**1.43**	47.56	25.47	96.1	8.9	0.0	21.3	**92.2**	22.0	40.1
	[23]	19.97	19.98	61.46	23.26	19.57	64.79	34.84	19.57	19.83	61.40	22.98	19.02	64.70	34.58	2.1	3.2	0.0	0.0	6.1	0.0	1.9
	Ours w/o NIR	1.27	62.83	22.18	16.15	**3.47**	25.77	21.95	0.72	54.41	22.06	12.01	1.78	25.49	19.41	**96.2**	11.6	0.0	31.1	88.7	0.0	37.9
NIR	[27]	16.59	14.86	14.53	17.93	18.08	18.09	16.68	10.17	9.58	9.42	11.85	13.83	13.84	11.45	51.1	52.9	51.2	44.7	28.1	28.1	42.7
	[12]	13.37	16.14	13.55	14.57	25.41	25.41	18.08	9.20	11.08	9.74	10.98	21.72	21.71	14.07	51.9	50.2	49.5	46.6	21.0	21.1	40.0
	Ours w/o RGB	1.72	1.22	1.83	2.45	9.58	10.06	4.48	1.02	**0.61**	**1.00**	1.34	6.83	7.21	3.00	93.5	**97.7**	**93.9**	94.0	58.1	56.7	82.3
RGB/ NIR	[30]	7.76	10.92	22.20	21.86	23.92	25.85	18.75	4.35	6.85	22.08	18.43	18.41	25.58	15.95	82.1	72.0	0.0	28.8	39.9	0.0	37.1
	[34]	1.30	2.26	2.56	**2.04**	9.57	12.39	5.02	0.75	1.27	1.41	**1.25**	6.57	9.38	3.44	95.8	91.1	91.1	**94.1**	68.9	47.7	81.5
	Ours	**1.26**	**1.21**	**1.82**	2.57	4.13	**9.75**	**3.46**	0.72	**0.61**	**1.00**	1.38	2.13	**6.95**	**2.13**	**96.2**	**97.7**	**93.9**	93.5	86.1	**57.2**	**87.4**

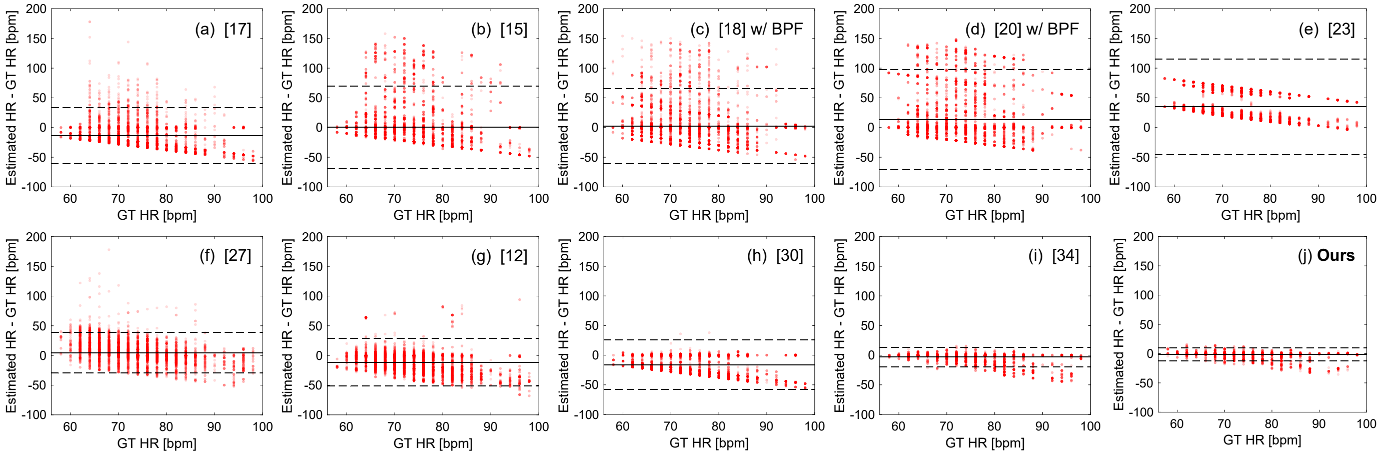


Fig. 6. Quantitative comparisons using Bland-Altman plots for all scenes and subjects in our dataset. In each figure, solid line represents the mean error, and dashed lines indicate 95 % limits of agreement between the estimated and ground-truth HRs.

conditions. However, in the scenes with varying illumination (“S3,” “S4,” and “S6”), the estimation performance of the RGB methods deteriorated significantly. This is primarily because it is difficult for the RGB methods to suppress the influence of largely varying illuminations. It was also observed that the estimation performance of the RGB-based methods was considerably degraded in “S2” more than that in “S3,” “S4,” and “S6.” We reason that the RGB videos taken in “S2” (low-light scene (0.4 lx)) were heavily deteriorated with noise; therefore the RGB-based methods failed to extract the signals attributable to the HR. Notably, it was observed that the results obtained using [20] and “[18] w/o BPF” showed worse performance than the other RGB-based methods. This is primarily because they did not utilize a band-pass filter that removes the frequency components outside the range of the HR. As can be observed, the methods “[20] w/ BPF” and “[18] w/ BPF” improved the estimation performance compared to those without using the band-pass filter.

In contrast, the NIR-based methods, especially “Ours w/o RGB,” showed better performance compared with the RGB-based ones in the varying illumination scenes (“S3,” “S4,” and “S6”). This is primarily because the reflecting NIR light could be captured with minimal spectrum noise in the normal HR range, as shown in the spectrograms of the background illumination variations (Fig. 5). Thus, the NIR-based method could stably extract the signals relevant to the HR, thereby yielding accurate HR estimations for such scenes.

It was observed that the method proposed in our earlier study [34] also outperformed the existing methods. However, in scenes where the participants moved considerably (“S5” and “S6”), the earlier method experienced performance degradation, unlike our current method. Conversely, the proposed method yielded accurate HR estimation under diverse illumination conditions as well as in scenes where the participants moved expansively.

2) Results for Other Datasets: The comparison results on the other datasets, “Tokyo” [37], “MR” [12], and “UBFC” [38], are listed in Table IV. Each value was obtained by averaging the results of all the subjects in each dataset. We find that our method can produce results comparable to those obtained using several state-of-the-art methods [18], [20]. This is primarily because these methods [18], [20] are designed to perform accurate HR estimation in controlled situations, such as scenes with minimal illumination variation. In practice, the datasets [37], [12], [38] did not include several sequences wherein there were significant illumination variations, as shown in the spectrograms of background illumination variations (Figs. 5 (g), (h), and (i)). Nevertheless, our method exhibited performance comparable with these methods. Hence, these comparison results suggest that our method enables accurate HR estimation not only in challenging situations where illumination significantly varies and the subjects moved expansively (our dataset) but also in controlled situations, as in the datasets [12], [37], [38].

We also find that the results obtained using our method were close to those obtained using "Ours w/o NIR" for "Tokyo" and "MR." As mentioned earlier, RGB signals are more reliable than NIR because of the light absorption characteristics of blood in stable and bright illumination scenes, as in these datasets. For this reason, we consider that our RGB/NIR fusion scheme adopted the estimates obtained using the RGB signals more than the NIR signals. It indicates that our method worked in the similar manner with "Ours w/o NIR."

The proposed method outperformed the deep learning-based methods [23], [26], [40], which are prone to overfitting the training datasets, as indicated by the comparison results. Thus, we infer that they degraded the accuracy of the HR estimation in other situations.

3) Comparisons in Time-Series Variations: We provide examples of the comparison results for the time-series variations in the HR estimation in Fig. 7. As can be observed in the middle row in Fig. 7, our method produced temporally stable and accurate outcomes, compared with the other methods.

The bottom row in Fig. 7 presents the time-series variations in the RGB and NIR confidence scores w^{RGB} and w^{NIR} obtained through the face/background correlation analysis. As can be observed, our method stably estimated that RGB observations were more reliable than NIR ones in the bright stable scenes ((a) "S1," (e) "S5," (g) "Tokyo," (h) "MR," and (i) "UBFC"). Conversely, the higher (more confidence) value for w^{NIR} than that for w^{RGB} could be obtained in the low-light ((b) "S2") and the varying illumination scenes ((c) "S3," (d) "S4," and (f) "S6"). These results yield that our RGB/NIR fusion scheme worked well.

E. Analysis

We analyzed the performance of our method more comprehensively. The impacts of the following contributions were first investigated: (1) adaptive fusion of RGB/NIR signals and (2) motion-robust time-series filtering. Thereafter, We evaluated the dependence of the HR estimation accuracy on varying κ, which is the specific hyperparameter required for our method. We also investigated the influence of the demosaicing methods for HR estimation performance. Finally, we discuss the limitations of our current method.

1) Impact of Adaptive Fusion of RGB/NIR Observations: We analyzed the effects of the adaptive fusion of the RGB/NIR observations. To spotlight these effects, we performed a maximum likelihood (ML) estimation of the latent HR h_τ. Specifically, we set the prior probability (i.e., past posterior probability) $p(h_\tau \,|\, Z_{1:\tau-1}) = 1$ and $w^{\mathrm{obs}}_\tau = 1$ in our entire framework. We refer to it as "Ours (ML)." We compared "Ours (ML)" with a method that equally exploited the RGB and NIR observations. Specifically, we set the weights such that $w^{\mathrm{RGB}}_\tau = w^{\mathrm{NIR}}_\tau = 1$ in the likelihood computation (Eq. (9)). We refer to it as "Ours (ML) w/o weights." Note that for the evaluation of the "UBFC," we set $w^{\mathrm{RGB}}_\tau = 1$ and $w^{\mathrm{NIR}}_\tau = 0$ because NIR videos were not included in this dataset.

The comparison results are presented in the first and second rows of Table V. It was observed that the ML-based method ("Ours (ML)") outperformed "Ours (ML) w/o weights." The impacts of the proposed RGB/NIR fusion scheme were most visible in the results for "S3" and "S6" in our dataset. As described earlier, the background illumination was sinusoidally oscillated at a frequency that was approximately equal to the normal HR in "S3" and "S6." In such challenging scenes, it is difficult to extract skin color changes related to HR from an RGB video. Thus, we infer that the method "Ours (ML) w/o weights" failed to achieve accurate HR estimation because of the equal incorporation of RGB and NIR observations.

2) Impact of Motion-Robust Time-Series Filtering: We investigated the effects of the motion-robust time-series filtering. In particular, we evaluated our method without motion analysis. Accordingly, we set the weight w^{obs}_τ in Eq. (26) as a constant value, $w^{\mathrm{obs}}_\tau = 0.5$. We refer to it as "Ours w/o motion."

The comparison results are presented in the third and fourth rows of Table V. As can be observed, compared with "Ours w/o motion," the proposed method ("Ours") yielded better HR estimates on "S5" and "S6" on our dataset where the subjects moved expansively.

We conducted additional experiments to clearly observe the effects of motion-robust time-series filtering. We first collected the subsequences in which the movement of the subject was significant relative to the other subsequences in each dataset. To do this, we introduce a parameter th that determines whether the subsequence includes the significant head movement of the subject. If the motion amount m_τ is larger than th, we consider that the corresponding subsequence is motion-significant. We collected the subsequences whose m_τ is larger than th. For these sequences, we compared our method with "Ours w/o motion." We used SR as a criterion to spotlight the impact of motion-robust time-series filtering. In this evaluation, we consider that the choice of th will vary the comparison results. Hence, we conducted the comparisons while varying th.

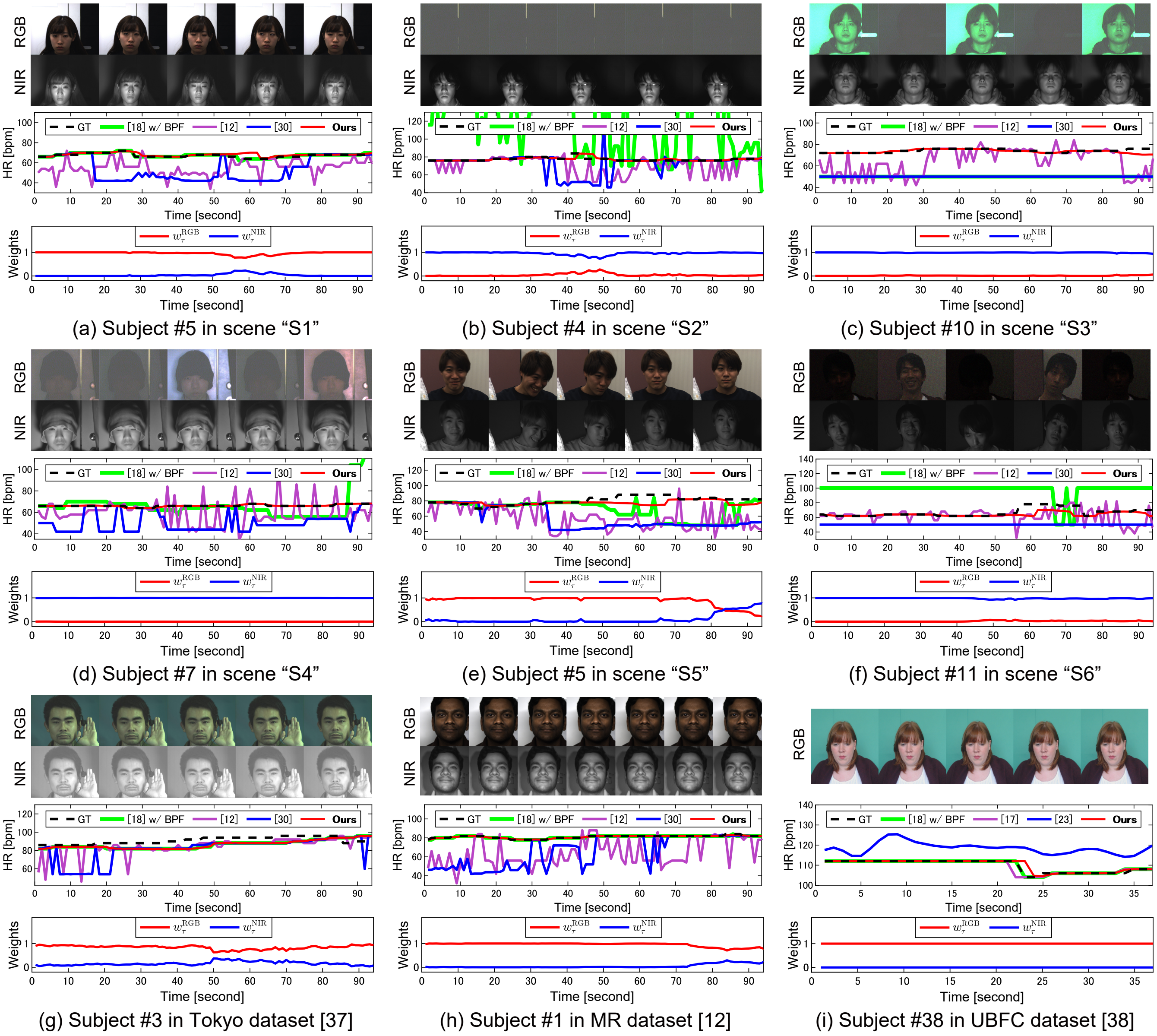
Figure 8 presents changes in the SR with varying th for each dataset. It was observed that the results obtained using "Ours" showed comparable with those obtained using "Ours w/o motion" for the datasets "S1," "S2," "S3," and "S4" as well as the other public datasets ("Tokyo," "MR," and "UBFC"). This is primarily because the sequences in these datasets contain less significant movements of the subjects, as shown in the motion histograms (Fig. 4). Thus, "Ours w/o motion" enabled the estimation of accurate HR outcomes for these datasets. Conversely, in the datasets "S5" and "S6" where the significant head movements are observed, "Ours" yielded the better performance than "Ours w/o motion." These results yield that motion-robust time-series filtering contributed to the improvement in robustness against motion artifacts in HR estimation.

Figure 9 presents an example of the comparisons in the time-series variations for the HR estimation in "S5." We also plotted the observation reliability w^{obs}_τ on the minor axis. It can be observed that our method produced temporally stable HR estimation results.

3) Influences of Varying κ: We analyzed the influence of the κ value, which is the hyperparameter of our Bayesian inference, on the HR estimation. We varied it such that $2 \leq \kappa \leq 20$.

TABLE IV
QUANTITATIVE COMPARISONS USING OTHER DATASETS [37], [12], AND [38] USING THE AVERAGE RMSE, MAE, AND SR. NOTE THAT WE LISTED THE RESULTS FOR [26], [40] ON UBFC DATASET, AS REPORTED IN RESPECTIVE PAPERS.

		RMSE (bpm)				MAE (bpm)				SR (%)			
		Tokyo [37]	MR [12]	UBFC [38]	Avg.	Tokyo [37]	MR [12]	UBFC [38]	Avg.	Tokyo [37]	MR [12]	UBFC [38]	Avg.
RGB	[17]	11.82	6.81	23.04	13.89	6.77	3.49	17.77	9.34	74.9	88.2	63.6	75.6
	[15]	1.92	3.11	5.93	3.65	1.18	1.72	4.54	2.48	89.2	90.3	83.4	87.6
	[18] w/o BPF	4.26	0.77	4.88	3.30	1.73	0.19	4.04	1.99	86.7	**99.2**	85.8	90.6
	[18] w/ BPF	2.10	**0.70**	4.08	**2.29**	**1.15**	**0.17**	3.55	**1.62**	89.1	**99.2**	86.9	**91.7**
	[20]	2.13	0.77	18.87	7.26	1.19	0.20	16.98	6.13	88.5	99.1	85.5	91.1
	[20] w/ BPF	2.12	0.73	**4.06**	2.30	1.18	0.18	**3.54**	1.64	88.7	99.1	**87.0**	91.6
	[23]	15.23	21.87	10.60	15.90	14.28	21.58	9.94	15.27	23.4	1.5	35.1	20.0
	[40]	n/a	n/a	8.64	8.64	n/a	n/a	5.45	5.45	n/a	n/a	n/a	n/a
	[26]	n/a	n/a	7.42	7.42	n/a	n/a	5.97	5.97	n/a	n/a	n/a	n/a
	Ours w/o NIR	2.09	0.77	4.27	2.38	1.29	0.31	3.66	1.75	89.2	98.6	85.6	91.1
NIR	[27]	14.07	18.73	n/a	16.40	9.73	14.30	n/a	12.01	51.6	33.8	n/a	42.7
	[12]	10.07	15.82	n/a	12.94	5.56	10.64	n/a	8.10	73.1	50.4	n/a	61.7
	Ours w/o RGB	2.10	6.27	n/a	4.19	1.30	3.65	n/a	2.47	89.1	87.8	n/a	88.5
RGB/ NIR	[30]	12.35	6.63	22.41	13.80	7.08	3.38	17.42	9.22	73.9	87.7	64.1	75.2
	[34]	**1.85**	0.96	5.08	2.63	1.22	0.40	4.39	2.00	**90.0**	98.1	83.5	90.5
	Ours	2.13	0.78	4.27	2.39	1.29	0.32	3.66	1.76	89.2	98.5	85.6	91.1



(a) Subject #5 in scene "S1"

(b) Subject #4 in scene "S2"

(c) Subject #10 in scene "S3"

(d) Subject #7 in scene "S4"

(e) Subject #5 in scene "S5"

(f) Subject #11 in scene "S6"

(g) Subject #3 in Tokyo dataset [37]

(h) Subject #1 in MR dataset [12]

(i) Subject #38 in UBFC dataset [38]

Fig. 7. Comparison results for time-series variations in estimated HR. For each subfigure, the top row shows RGB and NIR video sequences (note that the UBFC dataset does not include NIR videos). The middle row represents time-series variations in estimated HRs; "GT" indicates ground-truth HR. The bottom row represents the time-series variations in w^{RGB} and w^{NIR} obtained by face/background correlation analysis.

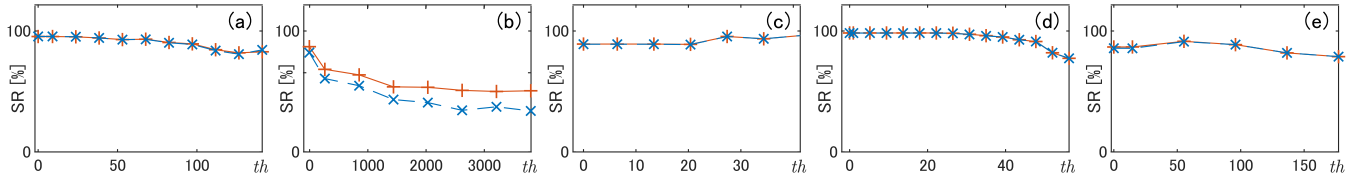


Fig. 8. Changes in success rate (SR) with varying th : (a) "S1," "S2," "S3," and "S4,"; (b) "S5" and "S6"; (c) Tokyo [37]; (d) MR [12]; (e) UBFC [38]. The orange and blue lines represent results obtained using "Ours" and "Ours w/o motion," respectively.

TABLE V

IMPACT OF ADAPTIVE FUSION OF RGB/NIR OBSERVATIONS (FIRST AND SECOND ROWS) AND MOTION-ROBUST TIME-SERIES FILTERING (THIRD AND FOURTH ROWS). WE EVALUATED PERFORMANCE USING THE RMSE.

	Our dataset						Tokyo	MR	UBFC	Avg.
	S1	S2	S3	S4	S5	S6	[37]	[12]	[38]	
Ours (ML) w/o weights	1.41	2.16	22.20	4.12	8.10	25.77	2.68	0.72	4.14	7.92
Ours (ML)	1.29	2.20	2.03	3.53	6.56	12.33	2.28	0.65	4.14	3.89
Ours w/o motion	1.24	1.19	1.81	2.62	5.08	11.12	2.13	0.79	4.35	3.37
Ours	1.26	1.21	1.82	2.57	4.13	9.75	2.13	0.78	4.27	3.10

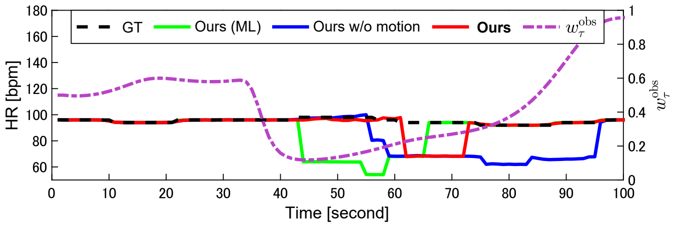


Fig. 9. Impact of motion-robust time-series filtering (#1 in "S5").

TABLE VI

CHANGES IN THE RMSE VALUES WITH VARYING κ.

	Our dataset						Tokyo	MR	UBFC	Avg.
	S1	S2	S3	S4	S5	S6	[37]	[12]	[38]	
$\kappa = 2$	1.31	0.96	1.85	2.14	3.61	9.72	2.22	0.98	4.82	3.07
$\kappa = 5$	1.26	1.21	1.82	2.57	4.13	9.75	2.13	0.78	4.27	3.10
$\kappa = 8$	1.23	1.38	1.83	2.82	4.21	9.97	2.01	0.72	4.16	3.15
$\kappa = 11$	1.27	1.69	1.87	3.03	4.48	10.08	2.06	0.68	4.12	3.25
$\kappa = 20$	1.26	2.22	1.96	3.32	4.88	10.60	2.14	0.64	4.01	3.45

TABLE VII

IMPACT OF DIFFERING DEMOSAICING METHODS. WE EVALUATED PERFORMANCE USING THE RMSE.

	S1	S2	S3	S4	S5	S6	Avg.
Ours (ML) w/ [48]	1.29	2.20	2.03	3.53	6.56	12.33	4.66
Ours (ML) w/ bicubic	1.28	2.19	2.03	3.53	6.42	12.33	4.63

The changes in the RMSE values with varying κ are listed in Table VI. The changes observed in the obtained RMSE values were insignificant, regardless of the κ value. Hence, the influence of κ on the HR estimation accuracy was minimal for the proposed method.

4) Influences of Demosaicing Methods: We also analyze the influence of differing demosaicing methods on HR estimation. As described earlier, the green component will include signals relevant to the HR more than the red and blue components. However, the demosaicing method [48] leverages the neighboring red and blue pixels to produce a full-resolution green channel, which may degrade the HR estimation performance.

To investigate the influences of the demosaicing process, we conducted additional experiments. We performed demosaicing of the RGB raw image using bicubic interpolation because it allows the production of each color channel with full resolution independently.

Table VII presents the comparison results using our dataset. We performed these comparisons in the ML estimation manner to spotlight the impacts of the different demosaicing methods. It can be observed that the results obtained with the different demosaicing methods were almost equivalent.

We discuss why the different demosaicing methods do not influence the performance of HR estimation. As described in Sect. III-A1, the peak spectral sensitivities of our imaging system for the blue, green, red, and NIR sensors appear at 460, 541, 608, and 797 nm, respectively. The FWHM of the spectral sensitivities for the blue, green, red, and NIR sensors are 90, 93, 92, and 151 nm, respectively. This indicates that the green pixels in the RGB raw image record a part of the red and blue components of the incident light. For this reason, we conjecture that our method with bicubic showed a performance similar to that of the demosaicing method [48], even though the red and blue channels were not utilized explicitly. From a different point of view, these comparisons imply that the effects of using the red and blue channels in *demosaicing of RGB raw images* are less significant for HR estimation.

Here, we discuss effects of demosaicing process on HR estimation more in-depth. In typical demosaicing methods, accurate color information can be produced in *non-textured* (flat) regions, whereas false colors and moiré, which are undesired image artifacts, are likely to be generated in textured regions. It implies that unreliable information may be obtained in textured regions. In contrast, our method extracts the time-series signals attributable to the HR from the facial region (forehead, cheek, or jaw) that contain *less-textured* information, thereby suggesting reliable signals with less demosaicing effects. Hence, we conjecture that the effects of the above-mentioned demosaicing problems would be less significant in HR estimation.

5) Limitation: In practice, we assumed that the subjects would move substantially over a short duration. Thus, when the subjects moved considerably for a long time, our method may fail to accurately estimate HR. To verify this limitation, we analyzed the performance of HR estimation in such cases.

An example of HR estimation results with low accuracy is shown in Fig. 10. In this case, the subject moved considerably for a long time, indicating that unstable observations for HR estimation were continuously obtained. We conjecture that our Bayesian inference fails to predict latent HR in this context.

To address this problem, we believe that it is important to compensate for the changes in the reflected light induced by the movement of a subject. We intend to investigate techniques to realize this in future research.

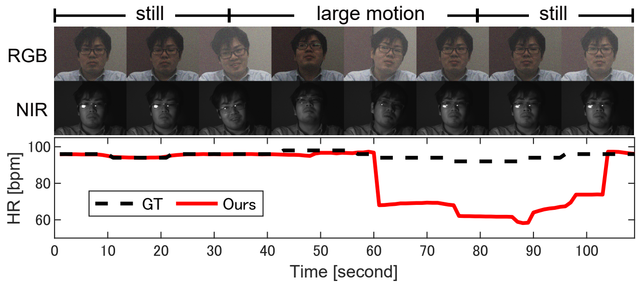


Fig. 10. Example of failure cases (#1 in "S6").

VIII. Conclusion

A. Summary

We proposed a non-contact HR estimation method that is robust to varying illumination conditions. We utilized an RGB/NIR camera to capture the temporal variations in the pixel value related to the cardiac pulse. We utilized RGB and NIR observations for HR estimation by analyzing the correlations between the signals in the face and background regions. To address the decrease in the HR estimation accuracy arising from the substantial movements of subjects, we incorporated a motion-robust time-series filtering into our framework. We demonstrated the effectiveness of the proposed method using public datasets and ours, including sequences with significant fluctuations in the illumination and movements of subjects.

B. Future Work

We intend to investigate techniques to address the limitations of our current method, as discussed in Section VII-E5.

We assumed that NIR face videos could be captured independently from varying background illumination. However, when varying background illumination includes significant NIR components, our method may fail to accurately estimate HR. In future work, we will investigate ways to address varying background illuminations, including multispectral information (such as ultraviolet and NIR).

References

[1] D. McDuff et al., "Remote measurement of cognitive stress via heart rate variability," in *Proc. of IEEE EMBC*, Aug. 2014, pp. 2957–2960.

[2] M. Burzo et al., "Towards sensing the influence of visual narratives on human affect," in *Proc. of ACM ICMI*, Oct. 2012, pp. 153–160.

[3] G. Valenza et al., "Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics," *Scientific Reports*, vol. 4, 2014.

[4] "Nevermind," https://nevermindgame.com/, Accessed: 2020-2-27.

[5] W. B. Fye, "A history of the origin, evolution, and impact of electrocardiography," *American Journal of Cardiology*, vol. 73, no. 13, pp. 937–949, 1994.

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

[7] X. Chen et al., "Video-based heart rate measurement: Recent advances and future prospects," *IEEE TIM*, vol. 68, no. 10, pp. 3600–3615, 2018.

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

[9] W. Verkruysse et al., "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[10] R. Stricker et al., "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. of IEEE RO-MAN*, Aug. 2014, pp. 1056–1062.

[11] F. Zhao et al., "Remote measurements of heart and respiration rates for telemedicine," *PLoS One*, vol. 8, no. 10, pp. 1–14, 2013.

[12] E. M. Nowara et al., "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Proc. of IEEE CVPRW*, Jun. 2018, pp. 1272–1281.

[13] M. Poh et al., "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE TBE*, vol. 58, no. 1, pp. 7–11, 2011.

[14] L. Wei et al., "Automatic webcam-based human heart rate measurements using laplacian eigenmap," in *Proc. of ACCV*, Nov. 2012, pp. 281–292.

[15] X. Li et al., "Remote heart rate measurement from face videos under realistic situations," in *Proc. of IEEE CVPR*, Jun. 2014, pp. 4264–4271.

[16] A. Subramaniam and K. Rajitha, "Spectral reflectance based heart rate measurement from facial video," in *Proc. of IEEE ICIP*, Sep. 2019, pp. 3362–3366.

[17] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. of IEEE ICCV*, Dec. 2015, pp. 3640–3648.

[18] S. Tulyakov et al., "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. of IEEE CVPR*, Jun. 2016, pp. 2396–2404.

[19] G. D. Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE TBE*, vol. 60, no. 10, pp. 2878–2886, 2013.

[20] W. Wang et al., "Algorithmic principles of remote ppg," *IEEE TBE*, vol. 64, no. 7, pp. 1479–1491, 2017.

[21] D. McDuff, "Deep super resolution for recovering physiological information from videos," in *Proc. of IEEE CVPRW*, Jun. 2018, pp. 1448–14 487.

[22] Y. Qiu et al., "Evm-cnn: Real-time contactless heart rate estimation from facial video," *IEEE TMM*, vol. 21, no. 7, pp. 1778–1787, 2019.

[23] R. Spetlík et al., "Visual heart rate estimation with convolutional neural network," in *Proc. of BMVC*, Sep. 2018, pp. 3–6.

[24] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. of ECCV*, Sep. 2018, pp. 349–365.

[25] Z. Yu et al., "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. of IEEE ICCV*, Oct. 2019, pp. 281–292.

[26] E. Lee et al., "Meta-rPPG: Remote heart rate estimation using a transductive meta-learner," in *Proc. of ECCV*, Aug. 2020, pp. 1–18.

[27] N. Martinez et al., "Non-contact photoplethysmogram and instantaneous heart rate estimation from infrared face video," in *Proc. of IEEE ICIP*, Sep. 2019, pp. 2020–2024.

[28] M. V. Gastel et al., "Motion robust remote-ppg in infrared," *IEEE TBE*, vol. 62, no. 5, pp. 1425–1433, 2015.

[29] S. B. Park et al., "Remote pulse rate measurement from near-infrared videos," *IEEE SPL*, vol. 25, no. 8, pp. 1271–1275, 2018.

[30] S. Kado et al., "Remote heart rate measurement from rgb-nir video based on spatial and spectral face patch selection," in *Proc. of IEEE EMBC*, Jul. 2018, pp. 5676–5680.

[31] W. Wang et al., "Discriminative signatures for remote-ppg," *IEEE TBE*, vol. 67, no. 5, pp. 1462–1473, 2020.

[32] L. F. C. Martinez et al., "Optimal wavelength selection for noncontact reflection photoplethysmography," in *Proc. of SPIE ICO*, vol. 8011, Nov. 2011, pp. 801 191–1–801191–7.

[33] E. B. Blackford et al., "Remote spectral measurements of the blood volume pulse with applications for imaging photoplethysmography," in *Proc. of SPIE BiOS*, vol. 10501, Feb. 2018, 105010Z–1–105010Z–8.

[34] K. Kurihara et al., "Adaptive fusion of RGB/NIR signals based on face/background cross-spectral analysis for heart rate estimation," in *Proc. of IEEE ICIP*, Sep. 2019, pp. 4534–4538.

[35] L. Feng et al., "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE TCSVT*, vol. 25, no. 5, pp. 879–891, 2015.

[36] C. Zhao et al., "Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography," in *Proc. of IEEE ICCVW*, Oct. 2019, pp. 1646–1655.

[37] Y. Maki et al., "Inter-beat interval estimation from facial video based on reliability of bvp signals," in *Proc. of IEEE EMBC*, Jul. 2019, pp. 6525–6528.

[38] S. Bobbia et al., "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, no. 1, pp. 82–90, 2019.

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

[40] F. Bousefsaf et al., “3D convolutional neural networks for remote pulse rate measurement and mapping from facial video,” *Applied Sciences*, vol. 9, no. 20, p. 4364, 2019.

[41] X. Niu et al., “Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video,” in *Proc. of ACCV*, Dec. 2018, pp. 562–576.

[42] X. Niu et al., “Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation,” *IEEE TIP*, vol. 29, pp. 2409–2423, 2020.

[43] L. Tarassenko et al., “Non-contact video-based vital sign monitoring using ambient light and auto-regressive models,” *Physiol. Meas.*, vol. 35, no. 5, pp. 807–831, 2014.

[44] D. Lee et al., “Heart rate estimation from facial photoplethysmography during dynamic illuminance changes,” in *Proc. of IEEE EMBC*, Aug. 2015, pp. 2758–2761.

[45] J. Cheng et al., “Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition,” *IEEE JBHI*, vol. 21, no. 5, pp. 1422–1433, 2017.

[46] L. Xu et al., “Illumination variation interference suppression in remote ppg using pls and memd,” *Electronics Letters*, vol. 53, no. 4, pp. 216–218, 2017.

[47] K. He et al., “Guided image filtering,” *IEEE TPAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.

[48] H. S. Malvar et al., “High-quality linear interpolation for demosaicing of bayer-patterned color images,” in *Proc. of IEEE ICASSP*, vol. 3, May 2004, pp. 485–488.

[49] M. S. Arulampalam et al., “A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking,” *IEEE TSP*, vol. 50, no. 2, pp. 174–188, 2002.

[50] Y. Fujita et al., “Parhelia: Particle filter-based heart rate estimation from photoplethysmographic signals during physical exercise,” *IEEE TBE*, vol. 65, no. 1, pp. 189–198, 2018.

[51] M. Kumar et al., “DistancePPG: Robust non-contact vital signs monitoring using a camera,” *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[52] E. Sánchez-Lozano et al., “A functional regression approach to facial landmark tracking,” *IEEE TPAMI*, vol. 40, no. 9, pp. 2037–2050, 2018.

[53] C. Yu et al., “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in *Proc. of ECCV*, Sep. 2018, pp. 325–341.

[54] L. Xi et al., “Image enhancement for remote photoplethysmography in a low-light environment,” in *Proc. of IEEE FG*, May 2020, pp. 485–491.

[55] P. Palatini, “Need for a revision of the normal limits of resting heart rate,” *Hypertension*, vol. 33, no. 2, pp. 622–625, 1999.

[56] D. Sugimura et al., “Enhancing color images of extremely low light scenes based on RGB/NIR images acquisition with different exposure times,” *IEEE TIP*, vol. 24, no. 11, pp. 3586–3597, 2015.

[57] T. Honda et al., “Low-light color image super-resolution using RGB/NIR sensor,” in *Proc. of IEEE ICIP*, Oct. 2018, pp. 56–60.

[58] T. Honda et al., “Multi-frame RGB/NIR imaging for low-light color image super-resolution,” *IEEE TCI*, vol. 6, pp. 248–262, 2020.

[59] H. Yamashita et al., “RGB-NIR imaging with exposure bracketing for joint denoising and deblurring of low-light color images,” in *Proc. of IEEE ICASSP*, Mar. 2017, pp. 6055–6059.

[60] H. Yamashita et al., “Low-light color image enhancement via iterative noise reduction using RGB/NIR sensor,” *Journal of Electronic Imaging*, no. 4, pp. 043 017–1–043017–13, 2017.

[61] H. Yamashita et al., “Enhancing low-light color images using an RGB-NIR sensor,” in *Proc. of IEEE VCIP*, Dec. 2015, pp. 1–4.

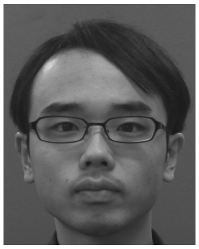
[62] T. Mikami et al., “Capturing color and near infrared images with different exposure times for image enhancement under extremely low-light scene,” in *Proc. of IEEE ICIP*, Oct. 2014, pp. 669–673.

[63] F. Andreotti et al., “Improved heart rate detection for camera-based photoplethysmography by means of kalman filtering,” in *Proc. of IEEE ICEN*, Apr. 2015, pp. 428–433.

[64] M. Poh et al., “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.

[65] J. M. Bland and D.Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” *Lancet*, vol. 327, no. 8476, pp. 307 –310, 1986.

[66] J. S. Krouwer, “Why bland-altman plots should use x, not (y+x)/2 when x is a reference method,” *Statistics in Medicine*, vol. 27, no. 5, pp. 778–780, 2008.



Kosuke Kurihara received the B.S. and M.S. degrees in electrical engineering from Tokyo University of Science, Tokyo, Japan, in 2018 and 2020, respectively. He is currently a Ph.D. student in Tokyo University of Science, Tokyo, Japan. His research interests include image processing.

Daisuke Sugimura (M’14) received the B.S. degree in engineering science from Osaka University, Osaka, Japan, in 2005, and the M.S. and Ph.D. degrees in information science and technology from the University of Tokyo, Tokyo, Japan, in 2007 and 2010, respectively. He is currently an associate professor with the Department of Computer Science, Tsuda University, Tokyo, Japan. His research interests include computer vision, machine learning and computational imaging.

Takayuki Hamamoto (S’95 - M’97) received the B.S. and M.S. degrees in electrical engineering from the Tokyo University of Science, Tokyo, Japan, in 1992 and 1994, respectively, and the Dr. Eng. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1997. He is currently a professor at the Department of Electrical Engineering at the Tokyo University of Science (TUS). His current research interests include image processing, computer vision and computational image sensors.