PHYSIOLOGICAL MODELING WITH MULTISPECTRAL IMAGING FOR HEART RATE ESTIMATION

Kosuke Kurihara[1], *Yoshihiro Maeda*[2], *Daisuke Sugimura*[3], *and Takayuki Hamamoto*[1]

[1]Tokyo University of Science [2]Shibaura Institute of Technology [3]Tokyo Metropolitan University

ABSTRACT

Heart rate (HR) is a key parameter in evaluating the physiological and emotional states of a person. In this paper, we propose a novel video-based heart rate (HR) estimation method based on physiological modeling with multispectral imaging. To capture blood volume pulse (BVP) associated with a person's heartbeat, we utilize a camera that records multispectral video consisting of red, green, blue, and near-infrared information. The novelty of the proposed method is the incorporation of a physiological BVP model into a multispectral HR estimation framework. The integration of a physiological model-based BVP signal extraction scheme into an adaptive multispectral framework enables the suppression of noise derived from ambient light and the accurate extraction of the BVP signal, thereby enhancing HR estimation performance. The experiments using RGB/NIR video datasets demonstrate the effectiveness of the proposed method.

Index Terms— Non-contact heart rate estimation, blood volume pulse, RGB/NIR camera

1. INTRODUCTION

Heart rate (HR) can provide insight into the physiological and emotional states of a person. In the last decade, non-contact video-based HR estimation methods have garnered considerable attention as alternatives to traditional contact-type HR measurement devices [1, 2]. Blood volume pulse (BVP) associated with cardiac pulse causes subtle temporal changes in skin color. The HR can be estimated by analyzing the temporal variation of reflected light from the skin, However, despite its significance, the performance of video-based HR estimation in the real world remains challenging.

This work was supported by JSPS KAKENHI Grant Numbers JP23KJ1961 and JP22K12080.

This article has been accepted for publication in IEEE International Conference on Image Processing (ICIP) 2024. 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/ICIP51287.2024.10647519

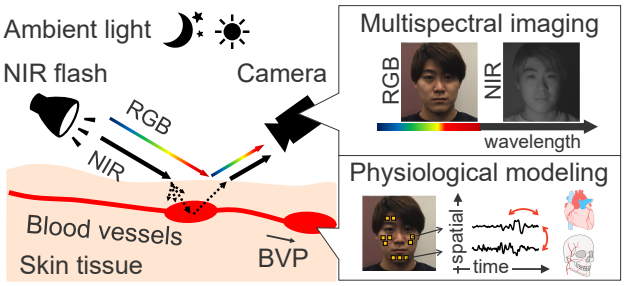

Fig. 1. Concept of proposed method.

The problem with video-based HR estimation is primarily the low signal-to-noise ratio (SNR) of the BVP signal. Two primary factors contribute to the low SNR of the BVP signal. First, as reported in [3], the temporal variation derived from BVP is notably small, typically less than 2 bits in an 8-bit depth video. Second, the camera records temporal variations of ambient illumination as well. Consequently, the recorded video encompasses both the subtle light derived from BVP and a significant amount of ambient light unrelated to the BVP signal, resulting in a low SNR of the BVP signal. Therefore, it is desirable to develop an HR estimation method that can effectively address these issues.

To address the first issue, researchers have developed methods based on the physiological modeling of the BVP phenomenon occurring within the human body, such as chrominance analysis based on the skin reflection model [4, 5] and spatio-temporal analysis based on the BVP physics model [6, 7]. Previous studies have shown that the performance of BVP signal extraction and the accuracy of the estimated HR can be improved by incorporating the physiological BVP model. In particular, spatio-temporal BVP modeling has been shown to be effective for accurate HR estimation. Kurihara *et al.* proposed a BVP signal extraction scheme that incorporates spatio-temporal BVP dynamical characteristics [6]. They modeled the periodic temporal characteristics of the BVP derived from the periodic cardiac pulse and the spatial similarity of the BVP across the face caused by propagation through facial blood vessels.

Multispectral imaging, typically near-infrared (NIR) imaging, has been demonstrated to mitigate the effects of ambient illumination variations on HR estimation [8, 9, 10, 11].

The influence of noise derived from uncontrolled illumination can be suppressed by using NIR video captured with the NIR flash unit. Furthermore, because NIR light is invisible to the human eye, it can be used without disturbing the surrounding environment. In the method [11], a framework was proposed that allows the flexible utilization of RGB and NIR observations for various illumination scenes.

Although these are effective in enhancing the performance of HR estimation, they encounter difficulties in simultaneously addressing both aforementioned issues.

In this study, we propose a novel method for HR estimation that leverages physiological modeling with multispectral imaging (Fig.1). The novelty of this study lies in the incorporation of the BVP model into the flexible RGB/NIR integration framework. The previous RGB/NIR method for HR estimation primarily addressed the noise derived from ambient illumination variations. By incorporating the BVP model into the RGB/NIR integration framework, we aim to enhance HR estimation performance.

2. PRELIMINARY

In this section, we briefly review the RGB/NIR integration method [11] and the spatio-temporal BVP signal extraction method, referred to as BVPDMD [6].

2.1. Flexible RGB/NIR Spectral Integration

In the method [11], HR estimation was performed based on Bayesian inference. 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 as N. We define all the observations up to the τ-th time window as $\mathbf{Z}_{1:\tau} = (\mathbf{z}_1, \ldots, \mathbf{z}_\tau)$.

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

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

where $P(\mathbf{z}_\tau \mid h_\tau)$, $P(h_\tau \mid h_{\tau-1})$, and $P(h_{\tau-1} \mid \mathbf{Z}_{1:\tau-1})$ denote the likelihood at τ, state transition probability from $\tau-1$ to τ, and posterior probability at $\tau-1$, respectively. In addition, $P(h_\tau \mid h_{\tau-1})$ is the state transition probability, which is modeled as a first-order autoregressive model because HR variations are expected to be small within a short duration [12, 13].

Among the terms in Eq. (1), the likelihood term $P(\mathbf{z}_\tau \mid h_\tau)$ plays a crucial role in determining the HR estimation performance. We enhance the performance of HR estimation by incorporating the BVP model when computing $P(\mathbf{z}_\tau \mid h_\tau)$. The details of this process are explained in Sect.3.3.

We calculate $P(h_\tau \mid \mathbf{Z}_{1:\tau})$ sequentially based on a particle filter framework [13]. Using the obtained $P(h_\tau \mid \mathbf{Z}_{1:\tau})$, we infer the latent HR h_τ^* based on the maximum a posteriori estimation (MAP) as

$$h_\tau^* = \arg\max_{h_\tau} \; P(h_\tau \mid \mathbf{Z}_{1:\tau}) \,. \tag{2}$$

2.2. Spatio-temporal BVP Signal Extraction

2.2.1. Naive DMD

DMD is a data-driven modal decomposition method based on a linear dynamical model. Suppose an observed discrete time-series signal $\mathbf{Y}_{1:e} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_e] \in \mathbb{R}^{l \times e}$ with e time length, where each element $\mathbf{y}_k$ has a l-dimensional component. DMD models the spatio-temporal dynamics of $\mathbf{Y}_{1:e}$ based on the following linear discrete dynamical system:

$$\mathbf{Y}_{2:e} \approx \mathbf{F}\mathbf{Y}_{1:e-1} \,, \tag{3}$$

where $\mathbf{F}$ denotes a state-transition matrix characterized by a linear dynamical system obtained by solving Eq. (3).

DMD seeks the dominant spatio-temporal dynamics by applying eigendecomposition to $\mathbf{F}$. The k-th time step signal modeled by a linear discrete dynamical system, that is, $\mathbf{y}_k \approx \mathbf{F}^{k-1}\mathbf{y}_1$, can be represented as

$$\mathbf{y}_k \approx \mathbf{F}^{k-1}\mathbf{y}_1 = \mathbf{\Psi}\mathbf{\Lambda}^{k-1}\mathbf{b} = \sum_{j=1}^{J} b_j \boldsymbol{\psi}_j \lambda_j^{k-1} \,, \tag{4}$$

where J denotes the number of DMD modes. The matrix $\mathbf{\Psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_J) \in \mathbb{C}^{l \times J}$ comprises J column vectors, where each vector $\boldsymbol{\psi}_j$ denotes the j-th eigenvector (referred to as DMD modes), and $\mathbf{\Lambda}$ denotes the corresponding DMD eigenvalues represented by the diagonal matrix form: $\mathbf{\Lambda} = \mathrm{diag}\ (\lambda_1, ..., \lambda_J)$. In addition, $\mathbf{b} = (b_1, b_2, ..., b_J)^\mathsf{T}$ represents the vector comprising the amplitude of each DMD mode. Based on the DMD theory, the j-th eigenvector $\boldsymbol{\psi}_j$ and the corresponding eigenvalues λ_j represent the spatial and temporal structures of the input signal, respectively.

2.2.2. BVPDMD With BVP Dynamics Modeling

BVPDMD incorporates the nonlinearity and quasi-periodicity of the BVP dynamics into a naive DMD framework. The following section briefly describes the BVPDMD framework. For further details, refer to the original study [6].

The direct application of DMD to extract the BVP signal containing nonlinear dynamical components is ineffective, because DMD assumes that the input signal follows a linear dynamical system. To address this issue, BVPDMD first transforms the observed signal extracted from multiple facial patches into a time-delay coordinate system. This enables the

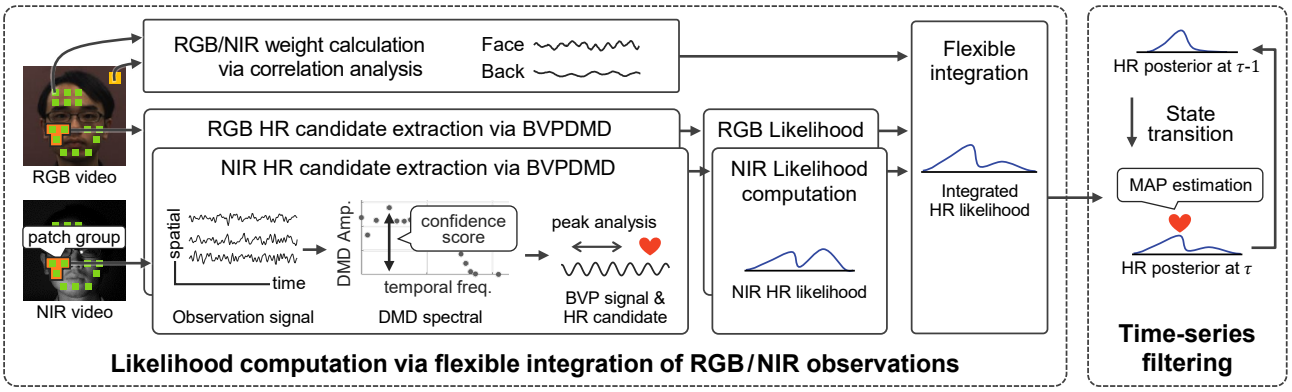

Fig. 2. Overview of the proposed method.

input signal to be approximated as a linear dynamical system, which can be analyzed using DMD.

To incorporate the quasi-periodic characteristics of BVP, BVPDMD models the BVP dynamics as an oscillating physical system. This is achieved through physics-informed DMD, which restricts the estimated state-transition matrix to be consistent with the prior knowledge on the physical structure.

BVPDMD then performs regularized optimization on the extracted DMD modes to promote the DMD modes in line with the spatio-temporal BVP characteristics.

To extract the BVP signal, BVPDMD identifies the DMD mode that is expected to contain the BVP components. BVPDMD selects the DMD mode with the highest DMD amplitude within the frequency band that is considered to be possessed by BVP, and thereafter extracts the BVP signal through an inverse time-delay transformation. Finally, HR can be estimated by applying a beat-to-beat peak period analysis to the extracted BVP signal.

3. PROPOSED METHOD

By incorporating physiological modeling of BVP into multi-spectral sensing, we enhance the performance of HR estimation. Specifically, we incorporate the BVPDMD framework into the likelihood computation based on flexible RGB/NIR integration (Fig.2). In the following section, we describe the proposed method in detail.

3.1. Obtaining Observation Signals From Facial Patches

First, image patches are selected from the RGB/NIR video. As in a previous study [11], we select patches and obtain their trajectory. We denote the set of the selected RGB and NIR patches by $\{\mathbf{P}_m^D\}_{m=1}^M$ $(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$, where M is the number of patches.

From $\{\mathbf{P}_m^D\}_{m=1}^M$ in the RGB and NIR video, we extract the RGB and NIR observation signals. The signals extracted from the RGB and NIR video in the τ-th time window are defined as $\{\mathbf{o}_{\tau,m}^c\}_{c\in\{\mathrm{R,G,B}\}}$, $\mathbf{o}_{\tau,m}^{\mathrm{NIR}}$, respectively, where the component $\mathbf{o}_{\tau,m}^c$ denotes a time-series signal in the $c(\in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\})$-th color channel.

3.2. Dividing Facial Patches Into Groups

To calculate the HR candidates, we perform BVPDMD for each group comprising adjacent patches, unlike in the original method in which BVPDMD was performed for all patches (i.e., the entire face). This is because, in an uncontrolled illumination scene, the observation signal is heavily deteriorated, and the spatio-temporal structure of the observation signals extracted from multiple patches becomes complex, making it difficult to analyze the observation signal extracted from the entire face at once. Because the spatio-temporal structures observed from adjacent patches are expected to be similar, analyzing them by each group makes the spatio-temporal analysis more tractable.

We first divide the facial patches into groups. The number of patches in each group is denoted as k. We group the patches such that the Euclidean distance between each patch in the same group is close, in an iterative manner. Let $\mathcal{G}_{m'}$ be the patch index set belonging to the m'-th group obtained in the m'-th iteration. We also define the patch index set that does not belong to any group for up to the m'-th iteration as $\mathcal{E}_{m'} = (\{1, ..., M\} \backslash \{\mathcal{G}_i\}_{i=1}^{m'})$, where $\mathcal{A}\backslash\mathcal{B}$ calculates the set of difference between two input sets $\mathcal{B}$ and $\mathcal{A}$. As an initial step, we set $\mathcal{E}_0$ to $\mathcal{E}_0 = \{1, ..., M\}$.

In the m'-th iteration, to form the m'-th group, we select the v-th patch $\mathbf{P}_v$ from $\mathcal{E}_{m'-1}$ and determine the k patches closest to $\mathbf{P}_v$ from $\mathcal{E}_{m'-1}\backslash v$. The set of patch indices belonging to the m'-th group $\mathcal{G}_{m'}$ is obtained as

$$\mathcal{G}_{m'} = \mathrm{Near}_k\left(\mathrm{Pos}(\mathbf{P}_v)\ ,\ \{\mathrm{Pos}(\mathbf{P}_w)\}_{w\in\mathcal{E}_{m'-1}\backslash v}\right), \quad (5)$$

where $\mathrm{Near}_k(\mathbf{h}, \mathcal{H})$ is an operator that determines the k-nearest neighbor patches to the input patch position $\mathbf{h}$ among the set of query patch positions $\mathcal{H}$ and returns the corresponding indices. The operator $\mathrm{Pos}(\cdot)$ returns the center position of the input patch.

This procedure is repeated until the M'-th iteration in which all patches belong to a group (i.e., $\mathcal{E}_{M'} = \varnothing$, where $\varnothing$ denotes an empty set), and we obtain $\{\mathcal{G}_{m'}\}_{m'=1}^{M'}$.

3.3. RGB/NIR Likelihood Computation using BVPDMD

Here, we describe the scheme for the computation of the likelihood of the RGB and NIR observations. We calculate the HR candidates from the RGB and NIR video by applying BVPDMD to each local patch set. We then compute the likelihood of RGB and NIR observation by aggregating all HR candidates.

3.3.1. HR Candidates Extraction From RGB Video

We transform the RGB signals into a color-difference space, which is effective for RGB video-based HR estimation [5, 4]. Based on a previous method [4], we project $\mathbf{o}_{\tau,m}^{c}$ onto the chrominance space and obtain the chrominance signal $\mathbf{u}_{\tau,m}$.

We calculate the HR candidates using BVPDMD [6]. For the m'-th RGB HR candidate, we apply BVPDMD to the chrominance components belonging to the m'-th group patch set $\mathcal{G}_{m'}$. Furthermore, to assess the reliability of the calculated HR candidate, we use the DMD amplitude as a confidence score. When BVPDMD is performed, the amplitude of the extracted DMD mode representing the BVP components can be obtained. The larger the amplitude of the selected DMD mode, the stronger the BVP components that the selected DMD mode is expected to contain. Thus, we can consider the corresponding HR candidate to be reliable.

The m'-th RGB HR candidate and the corresponding reliability, denoted by $\alpha_{\tau,m'}^{\mathrm{RGB}}$ respectively, are calculated as

$$(c_{\tau,m'}^{\mathrm{RGB}},\ \alpha_{\tau,m'}^{\mathrm{RGB}}) = \mathrm{BVPDMD}(\{\mathbf{u}_{\tau,v}\}_{v\in\mathcal{G}_{m'}})\ , \qquad (6)$$

where $(c,\ \alpha) = \mathrm{BVPDMD}(\mathcal{H})$ denotes the BVPDMD operator that computes the HR candidate c and corresponding DMD amplitude α from the set of time-series signals $\mathcal{H}$.

We perform BVPDMD for all patch sets and obtain a pair of HR candidates and the corresponding confidence scores, denoted by $(c_{\tau,1}^{\mathrm{RGB}}, \ldots, c_{\tau,M'}^{\mathrm{RGB}})$ and $(\alpha_{\tau,1}^{\mathrm{RGB}}, \ldots, \alpha_{\tau,M'}^{\mathrm{RGB}})$, respectively.

3.3.2. HR Candidates Extraction From NIR Video

We describe the extraction of HR candidates from an NIR video using BVPDMD in detail. Similarly to the RGB process described in Sect.3.3.1, we apply BVPDMD analysis to each grouped set of NIR patches to obtain a pair of HR candidates and the corresponding confidence scores.

The m'-th NIR HR candidate and the corresponding reliability, denoted as $c_{\tau,m'}^{\mathrm{NIR}}$ and $\alpha_{\tau,m'}^{\mathrm{NIR}}$ respectively, are calculated as

$$(c_{\tau,m'}^{\mathrm{NIR}},\ \alpha_{\tau,m'}^{\mathrm{NIR}}) = \mathrm{BVPDMD}(\{\mathbf{o}_{\tau,m}^{\mathrm{NIR}}\}_{m\in\mathcal{G}_{m'}})\ . \qquad (7)$$

3.3.3. RGB and NIR Likelihood Computation

Similar to the method [11], we model the RGB and NIR likelihoods $P_D(\mathbf{I}_\tau^D \mid h_\tau)\,(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ using the weighted kernel density estimation. We compute $P_D(\mathbf{I}_\tau^D \mid h_\tau)\,(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ as

$$P_D(\mathbf{I}_\tau^D \mid h_\tau) = \frac{1}{\sum_{m'=1}^{M'} \alpha_{\tau,m'}^D} \frac{1}{M'W} \sum_{m'=1}^{M'} \alpha_{\tau,m'}^D K\left(\frac{h_\tau - c_{\tau,m'}^D}{W}\right)\ , \qquad (8)$$

where $K(\cdot)$ denotes a Gaussian kernel with a bandwidth W.

3.4. Flexible RGB/NIR Likelihood Integration

The likelihood $P(\mathbf{z}_\tau \mid h_\tau)$ in Eq. (1) is formed by flexibly integrating the RGB and NIR likelihoods using the RGB and NIR weights $w_\tau^D\ (D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ as

$$P(\mathbf{z}_\tau \mid h_\tau) = \frac{w_\tau^{\mathrm{RGB}}}{w_\tau^{\mathrm{RGB}} + w_\tau^{\mathrm{NIR}}} P_{\mathrm{RGB}}(\mathbf{I}_\tau^{\mathrm{RGB}} \mid h_\tau) + \frac{w_\tau^{\mathrm{NIR}}}{w_\tau^{\mathrm{RGB}} + w_\tau^{\mathrm{NIR}}} P_{\mathrm{NIR}}(\mathbf{I}_\tau^{\mathrm{NIR}} \mid h_\tau), \qquad (9)$$

where $P_D(\mathbf{I}_\tau^D \mid h_\tau)\,(D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ denotes the RGB and NIR likelihood.

The weights $w_\tau^D\ (D \in \{\mathrm{RGB}, \mathrm{NIR}\})$ are defined as

$$w_\tau^{\mathrm{RGB}} = (1 - \beta_\tau^{\mathrm{back}})(1 - \gamma_\tau), \qquad (10)$$

$$w_\tau^{\mathrm{NIR}} = \beta_\tau^{\mathrm{back}}\,\gamma_\tau, \qquad (11)$$

where $\beta_\tau^{\mathrm{back}}$ is computed through correlation analysis of the RGB signals observed in the face and background regions. When background light provides stable and sufficient illumination, we expect that the signals relevant to the HR will be distinguishable from the background illumination, thereby indicating weak correlations between the face and background signals. Conversely, when the background light significantly fluctuates, a dominant color cast occurs on the face regions owing to the varying background illumination, indicating a strong positive correlation between the face and background signals.

To calculate $\beta_\tau^{\mathrm{back}}$, we utilize the green channel of the RGB signal. This is primarily because the green component includes more signals relevant to BVP than the red and blue components, and thus significantly affects the accuracy of the BVP signal extraction [14].

Based on these assumptions, to examine the positive correlation, we calculate $\beta_\tau^{\mathrm{back}}$ as

$$\beta_\tau^{\mathrm{back}} = \max(\rho_\tau,\ 0)\ , \qquad (12)$$

where ρ_τ is the Pearson correlation coefficient between the facial and background signals.

The other weight γ_τ assesses which signals measured from the RGB and NIR videos are more reliable using confidence scores of the RGB and NIR HR candidates. When the confidence scores of the RGB HR candidate $\{\alpha^{\mathrm{RGB}}_{\tau,m'}\}$ are higher than those of the NIR $\{\alpha^{\mathrm{NIR}}_{\tau,m'}\}$, we consider the RGB HR candidates to be more reliable for HR estimation.

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

$$\gamma_\tau = \frac{\sum_{m' \in \Omega^{\mathrm{NIR}}_\tau} \alpha^{\mathrm{NIR}}_{\tau,m'}}{\sum_{m' \in \Omega^{\mathrm{RGB}}_\tau} \alpha^{\mathrm{RGB}}_{\tau,m'} + \sum_{m' \in \Omega^{\mathrm{NIR}}_\tau} \alpha^{\mathrm{NIR}}_{\tau,m'}}\ , \tag{13}$$

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

3.5. HR Estimation by MAP Estimation

We estimate the HR based on Bayesian inference, as described in Sect.2.1. By incorporating the computed likelihood $P(\mathbf{z}_\tau \mid h_\tau)$ into Eq. (1), we calculate the posterior probability $P(h_\tau \mid \mathbf{Z}_{1:\tau})$. Finally, we infer the τ-th HR h^*_τ based on MAP estimation, as shown in Eq. (2).

4. EXPERIMENTS

4.1. Experimental Settings

4.1.1. Dataset

To demonstrate the effectiveness of our method, we conducted experiments on the following datasets: TUS [11], TokyoTech Remote PPG [15], and MR-NIRP [16]. Table 1 summarizes the details of each dataset.

The TUS dataset was obtained under various illumination scenes. Specifically, (i) bright (illuminance: 600 lx), (ii) low-light (illuminance: 0.4 lx), (iii) varying illumination with a sinusoidal frequency (0.83 Hz) close to the HR of a healthy person (illuminance: 1 lx), and (iv) realistic varying illumination, such as a theater (illuminance: 1 lx). We refer to these conditions as "S1," "S2," "S3," and "S4," respectively. In addition, to evaluate in more realistic scenarios, in "S5" and "S6," the participants are allowed to freely move in front of the camera. The illumination conditions of "S5" and "S6" are similar to those of "S1" and "S3," respectively.

4.1.2. Comparison Methods

We compared our method with the following state-of-the-art methods based on non-deep learning: RGB video-based methods [4, 17, 3, 18, 7, 6], NIR video-based methods [19, 16], and RGB/NIR video-based methods [8, 11]. Furthermore, we compared our method with two deep learning-based methods: RGB video-based methods [20, 21] and RGB/NIR video-based method [9]. We used pre-trained models published by these authors for a fair comparison.

Based on preliminary experiments, we set the parameters for our methods to $T = 150\,(5\,\mathrm{s})$ and $W = 2$. The parameters required for BVPDMD used in our method for HR candidate extraction were the same as those presented in the original study [6].

Table 1. Details of datasets used in experiments.

	TUS [11]	Tokyo [15]	MR [16]
Camera	Two-plate RGB/NIR	Single-chip RGB/NIR	RGB and NIR
# Subjects	18	8	8
# Videos	66	8	8
Resolution	1296×966	640×480	640×640
Frame rate	30 fps	30 fps	30 fps
Duration	120 s	180 s	180 s
Illumination	Bright, Dark, Varying	Bright	Bright

Table 2. Quantitative comparisons in TUS dataset using average MAE [bpm]. The best scores are presented in **bold**.

Method	S1	S2	S3	S4	S5	S6	Avg.
[4]	10.72	154.35	22.98	57.23	23.64	139.92	68.14
[17]	51.98	131.73	20.23	30.59	34.09	25.26	48.98
[3]	5.18	157.34	28.51	35.20	13.89	73.91	52.34
[18]	4.13	87.05	23.14	26.94	6.26	25.60	28.85
[7]	3.61	126.63	23.04	33.44	9.99	25.60	37.05
[6]	3.04	104.23	22.90	28.57	6.27	25.80	31.80
[20]	12.08	13.78	17.98	14.34	12.23	21.47	15.31
[21]	19.53	74.04	27.89	53.78	27.31	28.35	38.48
[19]	27.58	24.24	26.03	28.06	25.13	25.13	26.03
[16]	45.41	46.37	59.69	22.64	57.31	57.31	48.12
[8]	5.59	13.47	22.34	17.74	15.60	25.59	16.72
[9]	40.34	39.53	30.01	37.96	38.25	37.40	37.25
[11]	23.99	26.94	24.97	20.63	26.06	18.25	23.47
Ours	**2.20**	**5.46**	**6.43**	**6.80**	**4.18**	**14.30**	**6.56**

4.1.3. Evaluation Metrics

The results were quantitatively evaluated using the mean absolute error (MAE). Furthermore, we assessed the HR estimation performance using the Bland–Altman analysis [22, 23], 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.

4.2. Comparison Results

Table 2 summarizes the comparison results using the MAE on the TUS dataset. We also summarize the comparison results using the MAE in Tokyo and the MR dataset in Table 3. Our

Table 3. Quantitative comparisons in Tokyo and MR datasets [15, 16] based on the average MAE [bpm]. The best scores are presented in **bold**.

Method	Tokyo [15]	MR [16]	Avg.
[4]	22.11	17.29	19.70
[17]	14.62	12.91	13.76
[3]	11.17	3.71	8.81
[18]	8.72	4.96	6.84
[7]	3.39	2.24	2.81
[6]	4.11	2.36	3.24
[20]	8.27	6.67	7.47
[21]	51.25	16.72	33.98
[19]	21.53	30.06	25.80
[16]	127.68	52.21	89.95
[8]	5.64	8.44	7.04
[9]	36.77	38.79	37.78
[11]	15.17	26.68	20.93
Ours	**2.33**	**1.35**	**1.84**

Table 4. Impact of employing spatio-temporal BVP modeling within RGB/NIR sensing using TUS dataset. We evaluated the performance using the SR [%].

Method	S1	S2	S3	S4	S5	S6	Avg.
RGB/NIR wo/BVP	15.7	15.2	12.4	21.8	14.4	20.9	16.7
RGB/NIR w/BVP	90.9	61.1	54.0	62.2	76.6	35.1	63.3

method exhibits superior performance compared to the other methods on all datasets.

Fig 3 shows the comparison results using Bland–Altman plots. We can clearly observe that our method produces a narrow distribution around zero, indicating that it is capable of accurate HR estimation.

4.3. Analysis

We examined the impact of incorporating BVP modeling into RGB/NIR likelihood computations. To highlight these effects, we estimated the HRs using only the likelihood of observation (i.e., maximum likelihood estimation). We compared our method with a method that computes the RGB and NIR HR candidates from each patch without using BVPDMD. Specifically, instead of BVPDMD, we used a bandpass filter (0.7 – 4 Hz) with a frequency band expected to contain BVP to extract BVP signals from each patch of the RGB and NIR videos. We then performed peak detection on each extracted BVP signal to extract the HR candidates. To obtain the confidence score used to calculate the likelihood, we calculated the ratio of the largest amplitude to the second-largest amplitude of the power spectrum, similar to the original flexible RGB/NIR integration framework. We denote the above method and the method that incorporates BVP modeling by "wo/BVP" and "w/BVP," respectively.

As evaluation metrics. we used success rate (SR), which

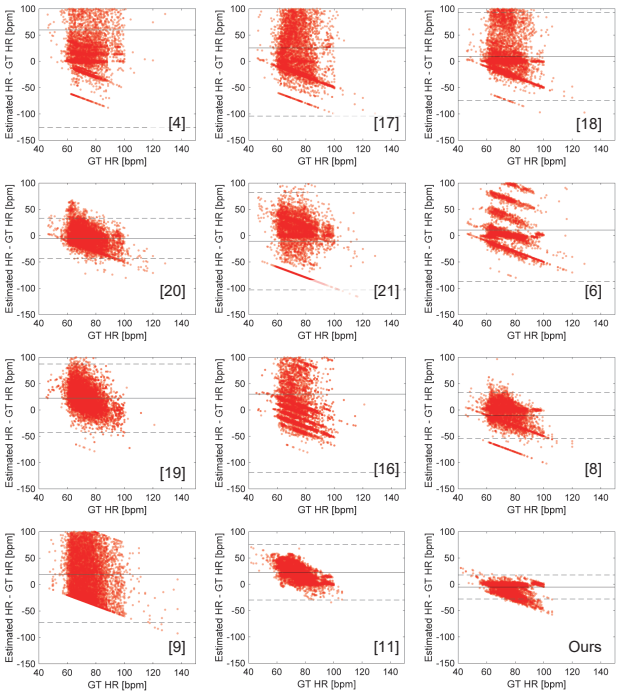

Fig. 3. Comparisons using Bland–Altman plots for all scenes and subjects in TUS dataset. The solid line represents the mean error, and the dashed lines indicate 95 % limits of agreement between the estimated and ground-truth HRs.

is the ratio of the number of successful HR estimation results to the total number of results. Following previous studies [8, 11, 6], we considered it successful if the HR estimation error was less than a certain threshold ($\pm$ 5 bpm).

Table 4 summarizes a comparison of the results for the TUS dataset. The method "w/BVP" outperformed the method "wo/BVP," indicating that BVP modeling contributes to the improvement of the HR estimation performance. In the method "wo/BVP," the BVP signal was extracted using a bandpass filter and frequency analysis, which is a general-purpose signal processing technique. Although the bandpass range was determined on the basis of the BVP characteristics, it is difficult to extract the BVP signal because the BVP signal is extremely weak and susceptible to noise. We infer that by incorporating the BVP model, the accuracy of the BVP signal extraction can be improved; thus, the accuracy of the HR estimation can also be improved.

5. CONCLUSION

We proposed a video-based HR estimation method based on physiological modeling with multispectral imaging. The HR estimation performance can be enhanced by incorporating a physiological BVP model into the RGB/NIR sensing scheme. We demonstrated the effectiveness of our method through experiments using RGB/NIR video datasets.

Despite these promising results, the proposed method faces challenges when implemented on devices with limited computational resources. The BVPDMD algorithm has high computational complexity because it iteratively solves optimization problems in a high-dimensional space [6]. We plan to investigate an algorithm that can solve problems with less complexity, such as first-order optimization.

6. REFERENCES

[1] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE TIM*, vol. 68, no. 10, pp. 3600–3615, 2018.

[2] 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.

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

[4] W. Wang, A. C. D. Brinker, S. Stuijk, and G. D. Haan, "Algorithmic principles of remote ppg," *IEEE TBE*, vol. 64, no. 7, pp. 1479–1491, 2017.

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

[6] Kosuke Kurihara, Yoshihiro Maeda, Daisuke Sugimura, and Takayuki Hamamoto, "Spatio-temporal structure extraction of blood volume pulse using dynamic mode decomposition for heart rate estimation," *IEEE Access*, vol. 11, pp. 59081–59096, 2023.

[7] Kosuke Kurihara, Yoshihiro Maeda, Daisuke Sugimura, and Takayuki Hamamoto, "Blood volume pulse signal extraction based on spatio-temporal low-rank approximation for heart rate estimation," in *Proc. of IEEE VCIP*, 2022, pp. 1–5.

[8] S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Remote heart rate measurement from rgb-nir video based on spatial and spectral face patch selection," in *Proc. of IEEE EMBC*, 7 2018, pp. 5676–5680.

[9] Lili Liu, Zhaoqiang Xia, Xiaobiao Zhang, Jinye Peng, Xiaoyi Feng, and Guoying Zhao, "Information-enhanced network for noncontact heart rate estimation from facial videos," *IEEE TCSVT*, 2023, Early Access.

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

[11] K. Kurihara, D. Sugimura, and T. Hamamoto, "Non-contact heart rate estimation via adaptive rgb/nir signal fusion," *IEEE TIP*, vol. 30, pp. 6528–6543, 2021.

[12] 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.

[13] Y. Fujita, M. Hiromoto, and T. Sato, "Parhelia: Particle filter-based heart rate estimation from photoplethysmographic signals during physical exercise," *IEEE TBE*, vol. 65, no. 1, pp. 189–198, 2018.

[14] L. F. C. Martinez, G. Paez, and M. Strojnik, "Optimal wavelength selection for noncontact reflection photoplethysmography," in *Proc. of SPIE ICO*, 11 2011, vol. 8011, pp. 801191–1–801191–7.

[15] Y. Maki, Y. Monno, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Inter-beat interval estimation from facial video based on reliability of bvp signals," in *Proc. of IEEE EMBC*, 2019, pp. 6525–6528.

[16] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Proc. of IEEE CVPRW*, 2018, pp. 1272–1281.

[17] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. of IEEE CVPR*, 6 2014, pp. 4264–4271.

[18] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. of IEEE CVPR*, 2016, pp. 2396–2404.

[19] N. Martinez, M. Bertran, G. Sapiro, and H. Wu, "Noncontact photoplethysmogram and instantaneous heart rate estimation from infrared face video," in *Proc. of IEEE ICIP*, 9 2019, pp. 2020–2024.

[20] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. of NeurIPS*, 2020, vol. 33, pp. 19400–19411.

[21] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proc. of IEEE CVPR*, 2022, pp. 4176–4186.

[22] 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.

[23] 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.