Received 22 May 2023, accepted 5 June 2023, date of publication 9 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284465

RESEARCH ARTICLE

Spatio-Temporal Structure Extraction of Blood Volume Pulse Using Dynamic Mode Decomposition for Heart Rate Estimation

KOSUKE KURIHARA[1], (Graduate Student Member, IEEE), YOSHIHIRO MAEDA[1], (Member, IEEE), DAISUKE SUGIMURA[2], (Member, IEEE), AND TAKAYUKI HAMAMOTO[1], (Member, IEEE)

[1]Department of Electrical Engineering, Tokyo University of Science, Tokyo 125-8585, Japan
[2]Department of Computer Science, Tsuda University, Tokyo 187-8577, Japan

Corresponding author: Kosuke Kurihara (4321701@ed.tus.ac.jp)

This work was supported by JSPS KAKENHI Grant Numbers 23KJ1961 and 22K12080.

ABSTRACT This article proposes a novel blood volume pulse (BVP) signal extraction method for heart rate (HR) estimation that incorporates medical knowledge of the spatio-temporal BVP dynamics. Previous methods merely exploited the spatial similarity of BVPs observed from multiple facial patches and performed the low-rank approximation to extract BVP signals. If noise components are superimposed over the entire face, the previous methods have difficulty distinguishing between the BVP component and noise even in the low-rank subspace. The main novelty of the proposed method is the exploitation of the BVP characteristics in the spatial and temporal domains in a unified manner based on a dynamic mode decomposition (DMD) framework, which is used to extract spatio-temporal structures from multidimensional time-series signals. To analyze the BVP dynamics that exhibit nonlinearity and quasi-periodicity, physics-informed DMD was performed on the time-series signals extracted from facial patches in a time-delay coordinate system. This approach enables the estimation of the DMD modes, which effectively represent the spatio-temporal structures of the BVP dynamics. The other novelty of the proposed method is the incorporation of medical knowledge of the HR frequency band to select the optimal DMD mode. By incorporating this medical knowledge of HR into the proposed framework, the proposed method can accurately estimate the BVP signal and HR. The experimental results obtained using three publicly available datasets yielded a root-mean-square error of the HR estimation results of 6.37 bpm, a 36.5 % improvement over the state-of-the-art methods.

INDEX TERMS Non-contact heart rate estimation, blood volume pulse, dynamic mode decomposition.

I. INTRODUCTION

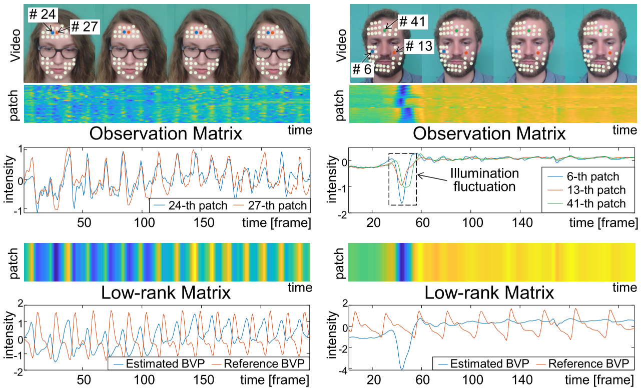
A. BACKGROUND AND OBJECTIVE

Heart rate (HR), defined as the number of cardiac pulses in a given period, provides insights into the physiological and emotional state of humans [1], [2], [3], [4]. HR can be measured by counting the number of cardiac pulses appearing within a certain time window.

Conventional HR measurement methods have used oximetry sensors that require physical contact with a subject [5], [6]. However, such contact-type sensors may cause discomfort or dermatitis to the subject [6]. Therefore, developing a method for non-contact HR estimation is desirable.

In the last decade, many researchers have proposed methods of non-contact HR estimation using cameras [6], [7], [8]. The blood volume pulse (BVP) associated with the cardiac pulse causes subtle temporal skin color changes in facial videos. By analyzing the temporal changes in skin color arising from the BVP, HR can be estimated. However, the

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.



(a) Stable illumination scene (b) Varying illumination scene

FIGURE 1. Problems of BVP signal extraction for HR estimation under varying illumination scenes. We constructed an observation matrix by stacking time-series signals from each patch region (white circular regions in each RGB facial video). We estimated the BVP signal using a low-rank approximation of the observation matrix in the spatial-domain. (a) In a stable illumination scene, a similar periodic characteristic attributable to BVP in each facial patch can be observed from the observation matrix and patch signal examples (second and third rows). Thus, the low-rank approximation in the spatial domain effectively works to extract the BVP signal. (b) In a varying illumination scene, many facial patches are subjected to illumination fluctuation (second and third rows), making it difficult to differentiate BVP and varying illumination components in the low-rank subspace in the spatial domain. This degrades the performance of the BVP signal extraction.

temporal variation of skin color owing to the cardiac pulse is quite small, which is less than 2 bits of the analog-to-digital converter of the camera [9]. Therefore, when noise is added to facial videos owing to the facial movements of the subject (e.g., facial expressions) and ambient illumination variations, the HR estimation performance is significantly degraded [9].

To overcome these problems, many methods have been proposed to exploit the spatial similarity of temporal changes in skin color attributable to the latent BVP signals on the face based on multiple facial patch representations [9], [10], [11], [12]. These methods assumed that the BVP signal can be observed similarly in neighboring facial patches because blood flows over the facial region at approximately the same time. Based on this assumption, they have claimed that the BVP signals can be represented in a low-rank subspace in the spatial domain. However, when many facial patches are subjected to similar noise due to movements of the subject or fluctuations in ambient illumination, accurate BVP signal extraction using these methods becomes difficult [9], [10], [11], [12]. This challenge arises primarily because such noise component exhibits similar characteristics throughout the facial region, indicating that they are also projected into the low-rank subspace in the spatial domain. Namely, even in a low-rank subspace, distinguishing between the BVP signal and noise is difficult.

Examples of BVP signals estimated using the low-rank approximation-based method [11] under a stable and varying illumination scene are shown in Fig. 1. In Fig. 1, (i) input RGB video sequences are presented in the first row, (ii) the second row shows the observation matrices for these RGB videos, (iii) the third row plots the time-series signals extracted from some of facial patches, (iv) the low-rank approximation results are presented in the fourth row, and (v) the last row presents the outcomes of the estimated and the corresponding reference HRs. In a stable scene ((a) in Fig. 1), periodic components arising from the BVP can be observed in every patch signal as well as in the entire observation matrix. Thus, the BVP signal can be accurately extracted using the spatial low-rank approximation. Conversely, the performance of BVP signal extraction is degraded in a varying illumination scene (Fig. 1 (b)). As mentioned earlier, the temporal color changes arising from the BVP are quite small; thus, accurate BVP signal extraction is difficult even when the illumination variation is small. Further, the entire face is subjected to illumination variation in this scene. Hence, similar characteristics owing to the noise component can be found in many facial patch signals and the observation matrix (second and third rows). Therefore, the noise component is dominant in the low-rank subspace in the spatial domain, making the spatial low-rank approximation method [11] inaccurate.

To address these problems, we exploit the BVP characteristics in the time domain. The medical field has widely agreed that the BVPs exhibit quasi-periodic characteristics in the time domain [13], [14]. As noise does not usually exhibit such a particular periodic behavior, if such quasi-periodic properties of the BVP can be extracted and exploited, the BVP component will be accurately distinguished from observations with noise artifacts.

In this study, we propose a novel method for BVP signal extraction based on spatio-temporal structure analysis using dynamic mode decomposition (DMD), a method for extracting the underlying spatio-temporal dynamics of a system from multi-dimensional time-series signals.

According to previous literature in the physics field [15], [16], [17], DMD operates well when the temporal behavior of the observed multi-dimensional signal can be modeled as a linear dynamical system. However, the direct application of DMD to the estimation of the BVP signal is ineffective because the propagation characteristics of the blood generally result in nonlinear temporal characteristics [18].

To address this problem, our method models the BVP dynamics exhibiting nonlinearity and quasi-periodicity, and incorporates them into the DMD framework, which is the main novelty of this study. According to previous research [15], [16], [17], a nonlinear signal can be modeled as a linear dynamical system in a time-delay coordinate system. Therefore, the BVP nonlinear dynamics can be modeled as linear dynamics in a time-delay coordinate system. Based on the quasi-periodicity of the BVP, we model the BVP dynamics as a conservative dynamical system, which is a physical system that oscillates without losing energy during the time range under consideration. Based on this modeling, we employ physics-informed DMD, a variant of DMD analysis that can incorporate the physical structure of conservative dynamical systems. The physics-informed DMD can restrict estimations that violate the physics law, making it less sensitive to noise. By performing physics-informed DMD in time-delay coordinates, the BVP component exhibiting nonlinearity and quasi-periodicity can be extracted from multidimensional time-series signals.

The other novelty of our method is the incorporation of medical knowledge of the frequency range of HR. Based on this knowledge, our method adaptively selects the optimal spatio-temporal structure that represents the latent BVP signal from those estimated by our physics-informed DMD-based framework.

By inverse time-delay embedding of the estimated best spatio-temporal structure, the BVP signal can be extracted. Finally, the HR is estimated from the extracted BVP signal by beat-to-beat peak period analysis.

The major contributions of this study can be summarized as follows.

- We propose a novel method for extracting the spatio-temporal structure of the BVP based on the physics-informed DMD in a time-delay coordinate system. This framework can model the BVP dynamics exhibiting nonlinearity and quasi-periodicity, enabling accurate BVP signal extraction that is less sensitive to noise.
- We present a scheme of adaptively selecting the plausible spatio-temporal structure attributable to the BVP signal from among those estimated with our physics-informed DMD framework based on the medical knowledge of the HR frequency range.

The remainder of this paper is organized as follows. Section I-C provides a comprehensive review of the related literature. Section I-D provides an overview of the proposed method. Section II describes the preprocessing of the input signals and the naïve DMD framework, as preliminaries. Section III presents the details of the proposed method. In particular, it describes the procedure for the observation matrix construction in time-delay coordinates and the details of the spatio-temporal structure analysis based on the physics-informed DMD framework. Section IV reports the experimental results obtained using publicly available datasets [10], [19], [20] to demonstrate the effectiveness of the proposed method. Finally, the conclusions and scope for future studies are provided in Section V.

B. NEW CONTRIBUTIONS TO OUR PREVIOUS STUDY

This study is an extension of our previous study, presented in [21]. Herein, we provide a brief explanation of [21] and present the new contributions of the present study.

In [21], the BVP signal was extracted by exploiting the spatial and temporal characteristics of the BVP in a hierarchical estimation manner. First, the BVP signal candidates are estimated by low-rank approximation in the time domain, where the quasi-periodic temporal behavior of the BVP is modeled using an autoregressive process. Then, the BVP signal is estimated by low-rank approximation in the spatial domain.

However, this approach has two limitations. First, the hierarchical estimation manner in [21] makes it difficult to exploit fully the spatio-temporal characteristics of BVP in estimating the BVP signal, which is primarily because the method [21] estimates the BVP candidate using an autoregressive model independently from each patch, indicating that the spatial characteristics of the BVP are discarded during this process. Thus, we consider that this process degrades the estimation accuracy of the BVP candidates, leading to performance degradation in the final BVP signal extraction. In contrast, this work ensures a unified spatio-temporal analysis, enabling the simultaneous exploitation of the spatio-temporal characteristics of the BVP.

The second limitation in [21] lies in the use of an autoregressive model to represent the quasi-periodic characteristics of the BVP. By definition, the autoregressive model expresses the temporal dependence of the latent BVP signal, as well as that of the noise components. Owing to this characteristic, BVP component and noise cannot easily be distinguished. In contrast, our current method can model the periodicity of the latent BVP signal based on a linear dynamical system in a time-delay coordinate system, enabling to distinction between BVP and noise components. To observe these claims experimentally, we compare our method with a hierarchical method [21] in Section IV-C2.

C. RELATED WORK

The framework for HR estimation from videos primarily comprises BVP signal extraction and HR estimation from the extracted BVP signal [7], [12], [13], [22], [23], [24], [25], [26], [27], [28], [29]. Once an accurate BVP signal has been extracted, an accurate HR outcome can be obtained. Thus, many researchers have focused on BVP signal extraction schemes for accurate HR estimation. The following section provides a comprehensive review of the related literature.

1) TEMPORAL SKIN-COLOR ANALYSIS

Verkruysse et al. [7] showed that the BVP signal can be extracted from temporal skin color changes arising from blood volume changes. They estimated the BVP signal using the green components extracted from the skin patches that were manually selected [7]. Poh et al. [30] modeled BVP signal extraction as a blind signal separation problem from the observed RGB signals. Haan and Jeanne [27] introduced novel chrominance features based on the analysis of a skin reflection model to eliminate specular reflection components that were unrelated to the HR.

2) LOW-RANK APPROXIMATION IN SPATIAL DOMAIN

Several studies have used the spatial characteristics of the BVP. BVPs observed from neighboring patches are similar because blood flows over the entire facial region at approximately the same time. Therefore, BVP signal estimation can be performed using multiple facial patch observations.

Kumar et al. [9] proposed an adaptive skin-patch selection scheme based on the weighted average of multiple patch observations. In this method [9], the weights for

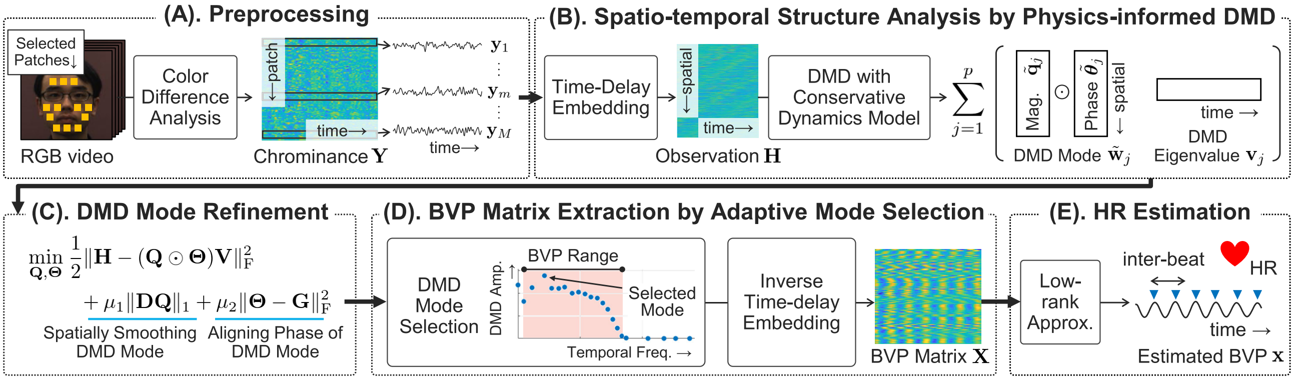

(A). Preprocessing
Selected Patches↓
RGB video
Color Difference Analysis
patch
time⟶
Chrominance $\mathbf{Y}$
$\mathbf{y}_1$
$\mathbf{y}_m$
$\mathbf{y}_M$
time⟶
(B). Spatio-temporal Structure Analysis by Physics-informed DMD
Time-Delay Embedding
spatial
time⟶
Observation $\mathbf{H}$
DMD with Conservative Dynamics Model
$\sum_{j=1}^{p}$
Mag. $\tilde{\mathbf{q}}_j$
$\odot$
Phase $\tilde{\boldsymbol{\theta}}_j$
spatial
DMD Mode $\tilde{\mathbf{w}}_j$
time ⟶
DMD Eigenvalue $\mathbf{v}_j$
(C). DMD Mode Refinement
$\min_{\mathbf{Q},\boldsymbol{\Theta}} \frac{1}{2}\|\mathbf{H}-(\mathbf{Q}\odot\boldsymbol{\Theta})\mathbf{V}\|_F^2 + \mu_1\|\mathbf{DQ}\|_1 + \mu_2\|\boldsymbol{\Theta}-\mathbf{G}\|_F^2$
Spatially Smoothing DMD Mode
Aligning Phase of DMD Mode
(D). BVP Matrix Extraction by Adaptive Mode Selection
DMD Mode Selection
BVP Range
DMD Amp.
Selected Mode
Temporal Freq. ⟶
Inverse Time-delay Embedding
BVP Matrix $\mathbf{X}$
(E). HR Estimation
Low-rank Approx.
inter-beat
HR
time ⟶
Estimated BVP $\mathbf{x}$

FIGURE 2. **Overview of our method: (A) Preprocessing for our spatio-temporal structure analysis is performed. The chrominance matrix is constructed using time-series signals extracted from multiple facial patches in an RGB facial video. (B) Spatio-temporal structure analysis of BVP is conducted by physics-informed DMD in a time-delay coordinate system with the assumption that the BVP dynamics approximately follow a conservative system. (C) DMD mode refinement is performed. Based on spatial characteristics of BVP over face region, regularized optimization is conducted to refine DMD modes. (D) BVP matrix extraction is conducted by inverse time-delay embedding of the outcome obtained using adaptive best DMD mode selection based on the medical knowledge of the frequency range of the BVP. (E) HR estimation in the time domain is performed by beat-to-beat peak period analysis of the extracted BVP signal.**

fusing multiple patch observations were determined based on the assumption that the dominant frequency component would be derived from the BVP signals if the patch signal contained components attributable to a cardiac pulse. Tulyakov et al. [11] introduced a matrix completion scheme for BVP signal extraction. They proposed a model in which a set of BVP candidates extracted from multiple patches could be represented in a spatial low-rank subspace. By using this model, they constructed an observation matrix using the observed patch signals and then performed a low-rank approximation on the observation matrix in the spatial domain. Nowara et al. [10] proposed a two-step method consisting of spatial low-rank approximation of multiple patch observations and noise reduction of the BVP candidates in the frequency domain. They assumed that the BVP candidates obtained from multiple facial patches contained sparse frequency components derived from the cardiac pulse. Based on this assumption, first, they performed a low-rank approximation on multiple patch observations in the spatial domain. Then, the BVP signal was extracted from the outcomes obtained in the first step based on the framework of the joint sparsity recovery in the frequency domain.

3) MACHINE-LEARNING APPROACH

In the last decade, many methods of BVP signal regression using a deep-learning framework have been proposed [31], [32], [33], [34], [35].

Niu et al. [31] constructed spatio-temporal maps using multiple facial patch signals. They fed the constructed maps into a regressor consisting of a convolutional neural network and a recurrent neural network. In the methods [32], [33], attention mechanisms were introduced for the regression of the BVP signal. Specifically, the authors proposed motion representation learning, which regresses the BVP signal using the difference in the patch signals between consecutive frames. They also introduced a facial appearance learning network based on attention mechanisms to facilitate motion representation learning. Yu et al. [34] proposed a transformer-based method for BVP signal regression using the global and local spatio-temporal features. The authors introduced a temporal difference transformer block to represent effectively the deep features attributable to the latent BVP signal.

4) ADVANTAGES OF PROPOSED METHOD

Herein, we summarize the primary differences between the aforementioned existing methods and the proposed approach.

In the methods [7], [30], accurate HR estimation is difficult when the face is subjected to noise components caused by lighting fluctuations or the movement of the face of the subject. This is primarily because skin color changes attributable to BVP are small, that is, less than 2 bit in an 8-bit RGB video [9]. To address the above-mentioned problems, methods based on low-rank approximation in the spatial domain [9], [10], [11] have been proposed using the spatial characteristics of the BVP. However, if many spatial patches are subjected to similar noise, distinguishing the noise component from the BVP is difficult because the noise component also satisfies the spatial similarity among neighboring patches.

In contrast to the above-mentioned existing methods, our method exploits the characteristics of the BVP in the time domain. Within the medical field, it is generally accepted that the BVP exhibits quasi-periodic characteristics in the time domain. As noise usually does not exhibit such a particular periodic behavior, if such quasi-periodic properties of the BVP can be extracted and exploited, the BVP component between noise artifacts will be accurately distinguished from the observations. To model the spatio-temporal characteristics of the BVP in a unified manner, our method utilizes a framework for spatio-temporal structure analysis of the BVP based on the physics-informed DMD framework.

```
Algorithm 1 Proposed HR Estimation Method
Input: RGB videos
1: Extract $M$ time-series signals from facial image patches using [21]
2: Construct chrominance matrix $\mathbf{Y}$ using [25]
3: Extract the DMD modes in the time-delay coordinate system using Algorithm 2
4: Perform DMD mode refinement based on regularized optimization using Algorithm 3
5: Extract the BVP matrix $\mathbf{X}$ by adaptive DMD mode selection using Algorithm 4
6: Estimate the HR $\mathcal{H}$ using beat-to-beat analysis
Output: Estimated HR $\mathcal{H}$
```

Methods using deep learning frameworks exhibit significant abilities to regress BVP signals [29], [31], [32], [33], [34]. Our method differs from these deep learning-based methods in that it relies on a model-based analysis. Further, we consider that there are considerable problems with these methods. In particular, we believe that they are prone to overfitting the training datasets when building their BVP regressors, suggesting that the accuracy of HR estimation would be reduced in other situations that are different from the training dataset. To observe this claim experimentally, we compared our method with some of these deep learning-based methods and report their results in Section IV.

D. OVERVIEW OF PROPOSED METHOD

An overview of the proposed method is presented in Fig. 2. First, we construct a chrominance matrix using time-series signals extracted from multiple facial patches in an RGB facial video ((A) in Fig. 2). Then, physics-informed DMD mode analysis in the time-delay coordinates is performed on the observation matrix ((B) and (C) in Fig. 2). Based on the medical knowledge of the frequency range of the BVP, the best DMD mode corresponding to the BVP signal is selected from among those estimated by the physics-informed DMD. The BVP signal can be extracted by inverse time-delay embedding of the estimated best DMD mode ((D) in Fig. 2). Finally, the HR is estimated by measuring the beat peak position of the estimated BVP signal ((E) in Fig. 2). We summarize the process of the proposed method comprehensively in Algorithm 1. The details of each subalgorithm used in the overall process are described in the subsequent sections.

II. PREPROCESSING AND PRELIMINARIES

This section describes the preprocessing procedure used to construct the chrominance matrix and briefly provides the fundamentals of the DMD framework.

A. CHROMINANCE MATRIX CONSTRUCTION

1) FACIAL PATCH SELECTION

Following the method [23], we first extract and track multiple facial patches over the input RGB video (the video length is denoted as N). By using the facial landmark positions obtained using the method [36], the entire face region is divided into M local patches in the first frame of the input RGB video. Then, we track each patch by performing a projective transformation based on the detected facial landmark positions between consecutive frames. By using the path tracking outcomes, M time-series RGB signals are obtained by averaging the pixel values within each patch. We denote a set of extracted RGB signals as $\mathbf{O} = \{\mathbf{o}_m\}_{m=1,2,\ldots,M}$, where $\mathbf{o}_m = \{\mathbf{o}_m^c\}_{c\in\{\mathrm{R,G,B}\}}$ represents the m-th RGB signal. For more details, refer to [23].

2) CHROMINANCE TRANSFORM

Next, we project $\mathbf{o}_m$ onto a color-difference space, primarily because analysis in a color-difference space enhances the BVP signal estimation performance [27].

Following the method [27], we first perform bandpass filtering (0.5–8 Hz) on $\mathbf{o}_m$ based on the knowledge of the HR range of a person [37]. For the m-th patch, we denote the bandpass-filtered RGB signal by $\mathbf{f}_m = \{\mathbf{f}_m^c\}_{c\in\{\mathrm{R,G,B}\}} = \mathrm{BPF}(\mathbf{o}_m)$, where $\mathrm{BPF}(\cdot)$ represents an operator for the bandpass filter of the input time-series signals. Then, we perform color-space conversion on the filtered signal $\mathbf{f}_m$ as

$$\mathbf{t}_m^{\mathrm{p1}} = 3\mathbf{f}_m^{\mathrm{R}} - 2\mathbf{f}_m^{\mathrm{G}}, \tag{1}$$

$$\mathbf{t}_m^{\mathrm{p2}} = 1.5\mathbf{f}_m^{\mathrm{R}} + \mathbf{f}_m^{\mathrm{G}} - 1.5\mathbf{f}_m^{\mathrm{B}}. \tag{2}$$

By using the projected components $\mathbf{t}_m^{\mathrm{p1}}$ and $\mathbf{t}_m^{\mathrm{p2}}$, the chrominance signal for the m-th patch, denoted as $\mathbf{y}_m = (y_m^{(1)}, \ldots, y_m^{(N)}) \in \mathbb{R}^{1\times N}$, is obtained as

$$\mathbf{y}_m = \mathbf{t}_m^{\mathrm{p1}} + \frac{\mathrm{Std}(\mathbf{t}_m^{\mathrm{p1}})}{\mathrm{Std}(\mathbf{t}_m^{\mathrm{p2}})}\,\mathbf{t}_m^{\mathrm{p2}}\,, \tag{3}$$

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

Finally, we construct the chrominance matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$ by stacking all chrominance signals $\{\mathbf{y}_m\}_{m=1}^{M}$ in the row dimension as

$$\mathbf{Y} = \begin{bmatrix} y_1^{(1)} & \cdots & y_1^{(N)} \\ & \vdots & \\ y_m^{(1)} & \cdots & y_m^{(N)} \\ & \vdots & \\ y_M^{(1)} & \cdots & y_M^{(N)} \end{bmatrix}. \tag{4}$$

B. FUNDAMENTALS OF DMD

Here, we provide a brief review of the fundamentals of DMD.

Suppose a discrete time-series signal $\mathbf{Z}_{1:e} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_e) \in \mathbb{R}^{l\times e}$ with e time length, where each element $\mathbf{z}_k$ has a l-dimensional component, that is, $\mathbf{z}_k \in \mathbb{R}^{l\times 1}$. DMD seeks the dominant dynamical component of $\mathbf{Z}_{1:e}$ based on the following linear discrete dynamical model:

$$\mathbf{Z}_{2:e} \approx \mathbf{A}\mathbf{Z}_{1:e-1}\,, \tag{5}$$

where $\mathbf{A}$ represents a state-transition matrix characterized by a linear dynamical system. Notably, the time evolution of a linear discrete dynamical system can be determined by applying $\mathbf{A}$ at each time step.

Mathematically, $\mathbf{A}$ is given by

$$\begin{aligned} \mathbf{A} &= \underset{\mathbf{A}}{\arg\min} \|\mathbf{Z}_{2:e} - \mathbf{A}\mathbf{Z}_{1:e-1}\|_{\mathrm{F}}^2 \qquad (6) \\ &= \mathbf{Z}_{2:e}(\mathbf{Z}_{1:e-1})^{\dagger} , \qquad (7) \end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm and $\dagger$ denotes the pseudo-inverse operator of the matrix.

The $(1+k)$-th time step signal modeled by a linear discrete dynamical system, i.e., $\mathbf{z}_{1+k} \approx \mathbf{A}^k\mathbf{z}_1$, can be represented by d DMD modes by applying eigendecomposition to $\mathbf{A}$:

$$\mathbf{z}_{1+k} \approx \mathbf{A}^k\mathbf{z}_1 = \boldsymbol{\Psi}\boldsymbol{\Lambda}^k\mathbf{b} = \sum_{j=1}^{d} b_j\boldsymbol{\psi}_j\lambda_j^k . \qquad (8)$$

In Eq. (8), the matrix $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_d) \in \mathbb{C}^{l\times d}$ comprises d column vectors, where each vector denotes the eigenvector (termed *normalized* DMD modes), and $\boldsymbol{\Lambda}$ denotes the corresponding DMD eigenvalues represented by the diagonal matrix form: $\boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d) \in \mathbb{C}^{d\times d}$. Additionally, $\mathbf{b} = (b_1, b_2, \ldots, b_d)^{\mathsf{T}} = \boldsymbol{\Psi}^{\dagger}\mathbf{z}_1 \in \mathbb{C}^{d\times 1}$ represents a vector comprising the amplitude of each DMD mode (termed DMD amplitude). Based on the DMD theory, the eigenvectors and eigenvalues, $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$, represent the spatio-temporal structures of the input signal, respectively. As shown in Eq. (8), the time evolution of each DMD mode from the 1st to the $(1+k)$-th time step can be represented by corresponding eigenvalue λ_j to the power of k.

Based on the above-described DMD mode decomposition procedure, $\mathbf{Z}_{1:e} = (\mathbf{z}_1, \ldots, \mathbf{z}_e)$ is decomposed into spectral components in the DMD basis. To obtain these outcomes, Eq. (8) can be reformulated as

$$\mathbf{z}_{1+k} \approx \boldsymbol{\Psi}\boldsymbol{\Lambda}^k\mathbf{b} = \boldsymbol{\Psi}\mathbf{B}\boldsymbol{\lambda}^k \equiv \boldsymbol{\Xi}\boldsymbol{\lambda}^k , \qquad (9)$$

where $\boldsymbol{\lambda}^k = (\lambda_1^k, \ldots, \lambda_d^k)^{\mathsf{T}}$ denotes the vector comprising the d eigenvalues, and $\mathbf{B}$ denotes the diagonal matrix form of $\mathbf{b}$, that is, $\mathbf{B} = \mathrm{diag}[\mathbf{b}]$.

By using the representation in Eq. (9), the outcome of the DMD of $\mathbf{Z}_{1:e}$ can be represented as

$$\begin{aligned} \mathbf{Z}_{1:e} &\approx \boldsymbol{\Xi} \begin{bmatrix} | & & | \\ \boldsymbol{\lambda}^0 & \cdots & \boldsymbol{\lambda}^{e-1} \\ | & & | \end{bmatrix} \\ &\equiv \boldsymbol{\Xi}\boldsymbol{\Gamma} \\ &= \underbrace{\begin{bmatrix} | & & | \\ \boldsymbol{\xi}_1 & \cdots & \boldsymbol{\xi}_d \\ | & & | \end{bmatrix}}_{\boldsymbol{\Xi}} \underbrace{\begin{bmatrix} — & \boldsymbol{\gamma}_1^{\mathsf{T}} & — \\ & \vdots & \\ — & \boldsymbol{\gamma}_d^{\mathsf{T}} & — \end{bmatrix}}_{\boldsymbol{\Gamma}} = \sum_{j=1}^{d} \boldsymbol{\xi}_j\boldsymbol{\gamma}_j^{\mathsf{T}} , \qquad (10) \end{aligned}$$

where $\boldsymbol{\xi}_j \in \mathbb{C}^{l\times 1}$ and $\boldsymbol{\gamma}_j \in \mathbb{C}^{e\times 1}$ represent the spatial and temporal structure of the j-th mode, respectively ($\boldsymbol{\xi}_j$ is termed DMD mode). Notably, $\boldsymbol{\gamma}_j$ characterizes the time evolution of the j-th DMD mode from the 1st to the e-th time step. It can be represented as $\boldsymbol{\gamma}_j = (\lambda_j^0, \ldots, \lambda_j^{e-1})^{\mathsf{T}}$.

By converting a dynamical system from a discrete domain into a continuous time domain with sampling frequency f_{s}, the temporal frequency of the j-th DMD mode can be obtained. In particular, Eq. (8) can be represented in the continuous time domain as

$$\mathbf{z}(t) \approx \sum_{j=1}^{d} b_j\boldsymbol{\psi}_j \exp(\omega_j t) = \sum_{j=1}^{d} b_j\boldsymbol{\psi}_j \exp(\alpha_j t)\exp(i\beta_j t) , \qquad (11)$$

where $i = \sqrt{-1}$ denotes the imaginary number, and t represents continuous time. In Eq. (11), $\omega_j = f_{\mathrm{s}} \ln(\lambda_j) \in \mathbb{C}$ characterizes the temporal behavior of the j-th DMD mode. Specifically, the real component of ω_j, denoted by α_j, represents the exponential growth and decay rate of the j-th mode, and the imaginary component of ω_j, denoted by β_j, represents the temporal frequency of the j-th mode.

III. SPATIO-TEMPORAL STRUCTURE EXTRACTION OF BVP DYNAMICS

This section details the primary novelty of this study, a scheme for extracting the spatio-temporal structure of BVP based on the physics-informed DMD that can incorporate the BVP dynamics exhibiting nonlinearity and quasi-periodicity.

A. MODELING BVP DYNAMICS

First, we discuss the modeling of the nonlinear and quasi-periodic dynamics of the BVP to apply our physics-informed DMD framework.

1) NONLINEAR BVP DYNAMICS

The spatio-temporal structures estimated using DMD provide the dominant dynamical behaviors of the input time-series signal. However, DMD assumes that the input time-series signals can be modeled using a linear dynamical system. Therefore, the direct application of DMD to the estimation of the BVP signal is ineffective because the propagation characteristics of the blood generally result in nonlinear temporal characteristics [18].

In contrast, some studies have reported that nonlinear signals can be modeled using a linear dynamical system in a time-delay coordinate system [15], [16], [17]. Based on these findings, we consider that the spatio-temporal structures of the BVP dynamics can be estimated by the DMD in the time-delay coordinate system.

2) QUASI-PERIODIC BVP DYNAMICS

As described earlier, the medical field has agreed that the BVP exhibits quasi-periodic dynamical behavior. We consider that the quasi-periodicity of BVP enables the BVP dynamics to be approximately modeled as an oscillating

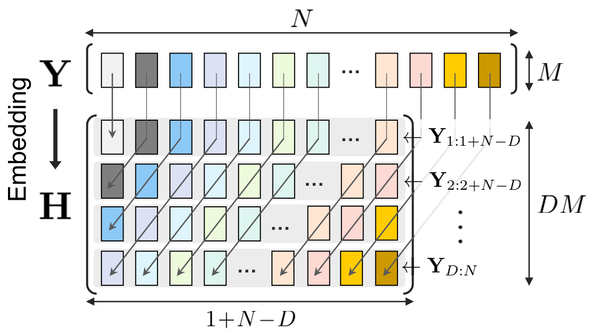


FIGURE 3. Illustration of time-delay embedding scheme.

physical system without exponential growth and decay amplitude components during the time range under consideration. On the other hand, in the literature on physics theory, dynamics modeled as a conservative system, in which the total amount of energy remains constant over time, exhibits periodic dynamical behaviors without exponential growth and decay amplitude components [38]. Therefore, the BVP dynamics can be approximately modeled as a conservative system.

Based on this modeling, we employ a physics-informed DMD, a variant of DMD analysis that can incorporate the physical structure of conservative dynamical systems. The physics-informed DMD can restrict estimations that violate the physics law, making it less sensitive to noise. On the basis of the above-mentioned knowledge, we estimate the spatio-temporal structure of BVP exhibiting quasi-periodically based on a physics-informed DMD framework.

B. DMD MODE ESTIMATION OF BVP IN TIME-DELAY COORDINATE SYSTEM

1) OBSERVATION MATRIX CONSTRUCTION BY TIME-DELAY EMBEDDING

To perform DMD analysis in the time-delay coordinate system, we construct observation matrices using time-delay embedding.

First, we compose the submatrices of $\mathbf{Y}$, defined as $\widehat{\mathbf{Y}} \in \mathbb{R}^{M\times N-1}$ and $\check{\mathbf{Y}} \in \mathbb{R}^{M\times N-1}$, by extracting from the 1st to the $(N-1)$-th, and from the 2nd to the N-th column vectors of $\mathbf{Y}$, respectively. Then, we embed $\mathbf{Y}$, $\widehat{\mathbf{Y}}$, and $\check{\mathbf{Y}}$ into the time-delay coordinate system, and use them as the input signals to our physics-informed DMD framework.

We describe the time-delay embedding procedure step by step as in [39]. For clarity, this process is illustrated in Fig. 3. We compose the submatrix of $\mathbf{Y}$ by extracting from the τ th to the $\tau+N-D$ th column vectors of $\mathbf{Y}$. We denote obtained submatrix as $\mathbf{Y}_{\tau:\tau+N-D}$, where D denotes the dimension parameter of the time-delay embedding. By performing above procedure while changing τ from 1 to D, we obtain a set of submatrices $\{\mathbf{Y}_{\tau:\tau+N-D}\}_{\tau=1}^{D}$. Then, we stack each of $\{\mathbf{Y}_{\tau:\tau+N-D}\}_{\tau=1}^{D}$ in the row dimension. This stacked matrix is defined as the observation matrix $\mathbf{H} \in \mathbb{R}^{DM\times(1+N-D)}$; it is represented as

$$\mathbf{H} = \begin{bmatrix} \mathbf{Y}_{1:1+N-D} \\ \vdots \\ \mathbf{Y}_{\tau:\tau+N-D} \\ \vdots \\ \mathbf{Y}_{D:N} \end{bmatrix} \equiv (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{1+N-D}) \,, \tag{12}$$

where $\mathbf{h}_\kappa \in \mathbb{R}^{DM\times 1}$ denotes the κ-th column vector of $\mathbf{H}$. We apply the above-described time-delay embedding procedure to $\widehat{\mathbf{Y}}$ and $\check{\mathbf{Y}}$, where the embedded submatrices are defined as $\widehat{\mathbf{H}} \in \mathbb{R}^{DM\times(N-D)}$ and $\check{\mathbf{H}} \in \mathbb{R}^{DM\times(N-D)}$, respectively.

Algorithm 2 DMD Mode Estimation in Time-Delay Coordinate System

```
Input: Chrominance matrix $\mathbf{Y}$
  Compute $\mathbf{H}$, $\widehat{\mathbf{H}}$, and $\check{\mathbf{H}}$ using Eq. (12)
  Compute $\mathbf{F}$ using Eq. (13)
  Compute $\widetilde{\mathbf{W}}$ and $\mathbf{V}$ by using eigendecomposition to $\mathbf{F}$
Output: DMD modes $\widetilde{\mathbf{W}}$ and DMD eigenvalues $\mathbf{V}$
```

2) PHYSICS-INFORMED DMD IN TIME-DELAY COORDINATE SYSTEM

By using the embedded matrices $\mathbf{H}$, $\widehat{\mathbf{H}}$, and $\check{\mathbf{H}}$, physics-informed DMD is performed to estimate the spatio-temporal structures of the BVP.

According to physics-informed DMD theory, a state-transition matrix in a conservative system can be modeled in a unitary matrix form [38]. Based on this finding, the state-transition matrix, denoted by $\mathbf{F}$, can be estimated as

$$\mathbf{F} = \underset{\mathbf{F}}{\arg\min} \, \|\check{\mathbf{H}} - \mathbf{F}\widehat{\mathbf{H}}\|_{\mathrm{F}}^{2} \quad \text{s.t.} \;\; \mathbf{F}\mathbf{F}^{*} = \mathbf{I} \,, \tag{13}$$

where $*$ denotes a Hermitian transpose operator, and $\mathbf{I}$ represents the identity matrix. In Eq. (13), the constraint, $\mathbf{F}\mathbf{F}^{*} = \mathbf{I}$, imposes that $\mathbf{F}$ must satisfy a unitary matrix form.

The outcome from the physics-informed DMD for the κ-th time step component, modeled as $\mathbf{h}_\kappa \approx \mathbf{F}^{\kappa-1}\mathbf{h}_1$, can be obtained by applying eigendecomposition to $\mathbf{F}$ as

$$\mathbf{h}_\kappa \approx \sum_{j=1}^{p} \tilde{\mathbf{w}}_j \eta_j^{\kappa-1} = \widetilde{\mathbf{W}} \boldsymbol{\eta}^{\kappa-1} \,, \tag{14}$$

where $\tilde{\mathbf{w}}_j \in \mathbb{C}^{DM\times 1}$ denotes the j-th DMD mode, and η_j denotes the corresponding j-th DMD eigenvalue. In addition, $\widetilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_p) \in \mathbb{C}^{DM\times p}$ represents the matrix comprising p DMD modes, and $\boldsymbol{\eta}^\kappa = (\eta_1^\kappa, \ldots, \eta_p^\kappa)^{\mathsf{T}} \in \mathbb{C}^{p\times 1}$ represents the vector form of the eigenvalues of the DMD modes at the κ-th time step.

In a matrix form, the outcome from the DMD of $\mathbf{H}$ can be obtained, resulting in the extraction of the DMD modes $\widetilde{\mathbf{W}}$ and the corresponding DMD eigenvalues $\mathbf{V}$. Mathematically, it can be represented as

$$\mathbf{H} = \underbrace{\begin{bmatrix} | & & | \\ \tilde{\mathbf{w}}_1 & \cdots & \tilde{\mathbf{w}}_p \\ | & & | \end{bmatrix}}_{\tilde{\mathbf{W}}} \underbrace{\begin{bmatrix} -\mathbf{v}_1^{\mathsf{T}}- \\ \vdots \\ -\mathbf{v}_p^{\mathsf{T}}- \end{bmatrix}}_{\mathbf{V}} = \sum_{j=1}^{p} \tilde{\mathbf{w}}_j \mathbf{v}_j^{\mathsf{T}} \,, \tag{15}$$

where $\mathbf{v}_j \in \mathbb{C}^{(1+N-D)\times 1}$ represents the vector comprising the state-transitioned eigenvalues of the j-th DMD eigenvalue: $\mathbf{v}_j = (\eta_j^0, \ldots, \eta_j^{N-D})^{\mathsf{T}}$. We summarize our DMD mode estimation scheme in Algorithm 2.

C. DMD MODE REFINEMENT BY REGULARIZED OPTIMIZATION

To suppress the noise components that still remain in the extracted DMD modes $\tilde{\mathbf{W}}$, we refine $\tilde{\mathbf{W}}$ based on an energy minimization process regularized with the above-mentioned spatial characteristics of the BVP. We denote the refined DMD mode to estimate as $\mathbf{W} \in \mathbb{C}^{DM\times p}$.

1) REPRESENTATION IN POLAR COORDINATE SYSTEM

To perform the DMD mode refinement, we first represent $\mathbf{W}$ in a polar coordinate system. In the polar coordinate system, the (u, v)-th component of $\mathbf{W}$, denoted by $[\mathbf{W}]_{u,v}$, can be represented as

$$[\mathbf{W}]_{u,v} = q_{u,v} \exp(i\theta_{u,v}) \,, \tag{16}$$

where $q_{u,v} \in \mathbb{R}$ and $\exp(i\theta_{u,v}) \in \mathbb{C}$ denote the magnitude and phase components of $[\mathbf{W}]_{u,v}$, respectively. In matrix form, $\mathbf{W}$ can be represented using an elementwise product as

$$\mathbf{W} = \mathbf{Q} \odot \boldsymbol{\Theta} \,, \tag{17}$$

where $\odot$ denotes the Hadamard product operator. The matrices $\mathbf{Q} = (q_{u,v})$ and $\boldsymbol{\Theta} = (\exp(i\theta_{u,v}))$ contain the magnitude and phase components of $\mathbf{W}$, respectively.

By using the magnitude and phase components of j-th DMD mode (i.e., the j-th column vectors of $\mathbf{Q}$ and $\boldsymbol{\Theta}$), respectively denoted as $\mathbf{q}_j$ and $\boldsymbol{\theta}_j$, the DMD outcome of the observation matrix $\mathbf{H}$ can be represented as

$$\mathbf{H} = (\mathbf{Q} \odot \boldsymbol{\Theta})\mathbf{V} = \sum_{j=1}^{p} (\mathbf{q}_j \odot \boldsymbol{\theta}_j)\, \mathbf{v}_j^{\mathsf{T}} \,. \tag{18}$$

2) REGULARIZED OPTIMIZATION

To obtain the refined magnitude and phase components of the DMD modes, we solve the following regularized optimization problem

$$\begin{aligned} \min_{\mathbf{Q},\boldsymbol{\Theta}} \ & \frac{1}{2}\|\mathbf{H} - (\mathbf{Q} \odot \boldsymbol{\Theta})\mathbf{V}\|_{\mathrm{F}}^2 \\ & + \mu_1 \|\mathbf{D}\mathbf{Q}\|_1 + \mu_2 \|\boldsymbol{\Theta} - \mathbf{G}\|_{\mathrm{F}}^2 \,, \end{aligned} \tag{19}$$

where the second and third terms represent the regularization terms, $\mathbf{D}$ denotes the matrix used for gradient computation in the spatial (i.e., column) direction of $\mathbf{Q}$, and $\|\cdot\|_1$ represents the L1 norm. In addition, μ_1 and μ_2 are control parameters for each regularization term.

We explain the operation of each regularization term to compensate for the spatio-temporal structures estimated in the DMD process. The first regularization term $\|\mathbf{D}\mathbf{Q}\|_1$ regularizes the spatial gradient of each mode $\mathbf{w}_j$ contained in $\mathbf{W}$. By using the L1 norm, our method regularizes the solution of Eq. (19) to be sparse, indicating that other noise components can be suppressed to nearly zero. Conversely, the second regularization term $\|\boldsymbol{\Theta} - \mathbf{G}\|_{\mathrm{F}}^2$ facilitates the alignment of the phase component of each DMD mode $\boldsymbol{\Theta}$, with the assistance of the guidance phase prior $\mathbf{G}$.

Because Eq. (19) can be formed as a block multiconvex function composed of multiple variables to be estimated (i.e., $\mathbf{Q}$ and $\boldsymbol{\Theta}$), we adopt an alternating iterative minimization procedure to solve this objective. Specifically, we numerically solve the following sub-problems alternately and iteratively. The solutions of $\mathbf{Q}$ and $\boldsymbol{\Theta}$ at the $(r + 1)$-th iteration, represented as $\mathbf{Q}^{(r+1)}$ and $\boldsymbol{\Theta}^{(r+1)}$, can be obtained as

$$\mathbf{Q}^{(r+1)} = \arg\min_{\mathbf{Q}} \frac{1}{2}\|\mathbf{H} - (\mathbf{Q} \odot \boldsymbol{\Theta}^{(r)})\mathbf{V}\|_{\mathrm{F}}^2 + \mu_1 \|\mathbf{D}\mathbf{Q}\|_1 \,, \tag{20}$$

$$\boldsymbol{\Theta}^{(r+1)} = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\|\mathbf{H} - (\mathbf{Q}^{(r+1)} \odot \boldsymbol{\Theta})\mathbf{V}\|_{\mathrm{F}}^2 + \mu_2 \|\boldsymbol{\Theta} - \mathbf{G}\|_{\mathrm{F}}^2 \,. \tag{21}$$

To solve the above convex problems (Eqs. (20) and (21)), we use the linear solver, the primal-dual interior-point (PDIP) method. As an initial solution of $\mathbf{Q}$ and $\boldsymbol{\Theta}$ (i.e., $\mathbf{Q}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$), we utilize $\tilde{\mathbf{Q}}$ and $\tilde{\boldsymbol{\Theta}}$ obtained from $\tilde{\mathbf{W}}$, respectively.

3) PHASE GUIDANCE CONSTRUCTION

Here, we describe a scheme for constructing $\mathbf{G} \in \mathbb{C}^{DM\times p}$. Generally, blood flows over the face at approximately the same time. This property suggests that the phase information of BVP waves tends to be aligned across all M facial patches. We exploit this property for the guidance construction for each DMD mode.

To construct the phase guidance for the j-th DMD mode $\boldsymbol{\theta}_j$, denoted by $\boldsymbol{g}_j \in \mathbb{C}^{DM\times 1}$, we calculate the average phase component of M patch signals of the initial solution of the j-th DMD mode used for the optimization (Eq. (19)). We denote the phase component of each patch signal as $\{\tilde{\boldsymbol{\theta}}_j^{(m)}\}_{m=1,2,\ldots,M}$. Notably, the averaging process enables noise reduction in the signal and thus facilitates reliable guidance construction for the phase alignment. Specifically, the average phase component of $\tilde{\boldsymbol{\theta}}_j$, denoted by ζ_j, is given by

$$\zeta_j = \frac{1}{M} \sum_{m=1}^{M} \tilde{\boldsymbol{\theta}}_j^{(m)} \,. \tag{22}$$

By using ζ_j, we compose $\boldsymbol{g}_j$. As described in Section III-B1, time-delay embedding involves creating a higher-dimensional space by shifting and stacking multiple copies of the M-dimensional facial patch time-series signals, with

Algorithm 3 DMD Mode Refinement

```
Input: DMD modes $\widetilde{\mathbf{W}}$
Compute $\widetilde{\mathbf{Q}}$ and $\widetilde{\boldsymbol{\Theta}}$ from $\widetilde{\mathbf{W}}$ using Eq. (17)
Compute $\mathbf{G}$ using Eqs. (22), (23), and (24)
Initialize:
    $r \leftarrow 0$
    $\mathbf{Q}^{(0)} \leftarrow \widetilde{\mathbf{Q}}$
    $\boldsymbol{\Theta}^{(0)} \leftarrow \widetilde{\boldsymbol{\Theta}}$
repeat
    Estimate $\mathbf{Q}^{(r+1)}$ by solving Eq. (20) with PDIP
    Estimate $\boldsymbol{\Theta}^{(r+1)}$ by solving Eq. (21) with PDIP
    $r \leftarrow r + 1$
until Convergence
Compute $\mathbf{W}$ using obtained $\mathbf{Q}$ and $\boldsymbol{\Theta}$ using Eq. (17)
Output: Refined DMD modes $\mathbf{W}$
```

each copy offset by a fixed time delay. Therefore, the phase component of each DMD mode $\boldsymbol{\theta}_j$ also has such shifted and stacking structure. Furthermore, in the j-th DMD modes, the time evolution (i.e., shift) can be expressed by multiplying η_j, as can be seen in Eq. (14). Based on such properties of DMD and the time-delay coordinate system, we construct $\boldsymbol{g}_j \in \mathbb{C}^{DM\times 1}$ as

$$\boldsymbol{g}_j \equiv \begin{bmatrix} \boldsymbol{\zeta}_j \\ \eta_j \boldsymbol{\zeta}_j \\ \vdots \\ \eta_j^D \boldsymbol{\zeta}_j \end{bmatrix} , \tag{23}$$

where $\boldsymbol{\zeta}_j$ is defined as $\boldsymbol{\zeta}_j = \zeta_j(1, 1, \dots, 1)^{\mathsf{T}} \in \mathbb{C}^{M\times 1}$.

We perform this processing for each DMD mode and then obtain the set $\{\boldsymbol{g}_j\}_{j=1}^{p}$. By using $\{\boldsymbol{g}_j\}_{j=1}^{p}$, the guidance phase prior $\mathbf{G} \in \mathbb{C}^{DM\times p}$ can be obtained as

$$\mathbf{G} = [\boldsymbol{g}_1 \;\; \boldsymbol{g}_2 \;\; \cdots \;\; \boldsymbol{g}_p] \,. \tag{24}$$

We summarize our DMD mode refinement scheme in Algorithm 3.

D. BVP MATRIX ESTIMATION BASED ON ADAPTIVE MODE SELECTION

We extract the BVP matrix $\mathbf{P}$, which is expected to contain the BVP signal to be estimated, from the spatio-temporal structures estimated using our framework. To this end, we need to determine the spatio-temporal structure from $\mathbf{W}$ and $\mathbf{V}$ that corresponds to $\mathbf{P}$.

1) ADAPTIVE MODE SELECTION FOR BVP DYNAMICS

Our method selects the best DMD mode that represents the spatio-temporal structures of BVP from $\mathbf{W}$ and $\mathbf{V}$. In this process, we consider that (i) the most-dominant spatio-temporal structure represents the BVP matrix $\mathbf{P}$ and (ii) $\mathbf{P}$ should be within the temporal frequency range of the BVP [37]. Based on these assumptions, we first calculate the j-th DMD amplitude χ_j and temporal frequency ρ_j as

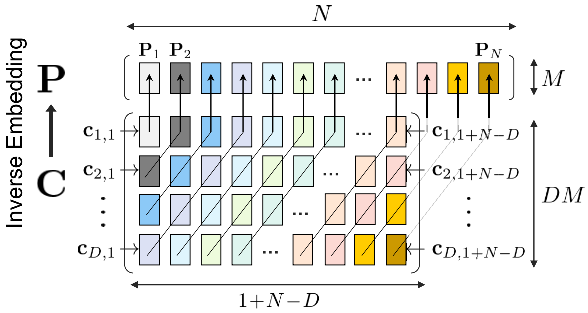


FIGURE 4. Illustration of inverse time-delay embedding scheme.

Algorithm 4 BVP Matrix Estimation

```
Input: DMD modes $\mathbf{W}$ and DMD eigenvalues $\mathbf{V}$
Compute $\{\chi_j\}_{j=1}^{p}$ and $\{\rho_j\}_{j=1}^{p}$ using Eqs. (25) and (26)
Determine $\varsigma$ based on Eq. (27)
Compute $\mathbf{C}$ from $\chi_\varsigma$ and $\rho_\varsigma$ using Eq. (28)
Compute $\mathbf{P}$ from $\mathbf{C}$ using Eqs. (29) and (30)
Output: BVP matrix $\mathbf{P}$
```

$$\chi_j = \|\mathbf{w}_j\|_2^2 \,, \tag{25}$$

$$\rho_j = f_{\mathrm{s}} \ln(\eta_j) \,, \tag{26}$$

where f_{s} denotes the frame rate of the video.

By using χ_j and ρ_j, we estimate the index number of the DMD mode that best represents $\mathbf{P}$, defined as ς, as

$$\varsigma = \arg\max_j \; (\chi_j) \quad \text{s.t.} \;\; \nu_{\mathrm{l}} < \rho_j < \nu_{\mathrm{h}} \,, \tag{27}$$

where ν_{l} and ν_{h} denote parameters determined based on the temporal frequency range of the BVP.

We extract the BVP component $\mathbf{C} \in \mathbb{R}^{DM\times(1+N-D)}$ using the selected ς-th DMD mode. Because the BVP component must comprise the real parts of the expanded components, $\mathbf{C}$ is obtained as

$$\mathbf{C} = \mathrm{Real}(\mathbf{w}_\varsigma \mathbf{v}_\varsigma^{\mathsf{T}}) \,, \tag{28}$$

where $\mathrm{Real}(\cdot)$ denotes an operator that extracts the real components of the input complex matrix.

2) INVERSE TIME-DELAY EMBEDDING

We perform inverse time-delay embedding on $\mathbf{C} \in \mathbb{C}^{DM\times(1+N-D)}$ to obtain the BVP matrix $\mathbf{P} \in \mathbb{R}^{M\times N}$. This procedure is illustrated in Fig. 4.

First, we represent $\mathbf{C}$ using M-dimensional column vectors $\{\mathbf{c}_{\varepsilon,\delta}\}_{\varepsilon=1,2,\dots,D,\; \delta=1,2,\dots,(1+N-D)}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{1,1} & \mathbf{c}_{1,2} & \cdots & \mathbf{c}_{1,1+N-D} \\ \mathbf{c}_{2,1} & \mathbf{c}_{2,2} & \cdots & \mathbf{c}_{2,1+N-D} \\ & \ddots & \ddots & \\ \mathbf{c}_{D,1} & \mathbf{c}_{D,2} & \cdots & \mathbf{c}_{D,1+N-D} \end{bmatrix} , \tag{29}$$

where $\mathbf{c}_{\varepsilon,\delta}$ is defined as $([\mathbf{C}]_{(\varepsilon-1)M+1,\delta}\ ,\ldots,\ [\mathbf{C}]_{\varepsilon M,\delta})^{\mathsf{T}}$ (i.e., the sub-vector of the δ-th column vector of $\mathbf{C}$ that is comprised of the row components ranging from the $(\varepsilon-1)M+1$ to εM-th ones).

We explain how the time evolution of the dynamics in the inverse time-delay (i.e., real space-time) coordinate system can be obtained from those represented in the time-delay coordinate system. Theoretically, the ε-th anti-diagonal components (i.e., $\mathbf{c}_{1,\varepsilon}, \mathbf{c}_{2,\varepsilon-1}, \ldots, \mathbf{c}_{\varepsilon-1,2}, \mathbf{c}_{\varepsilon,1}$) correspond to the M patch signals at the ε-th time step in the real space-time coordinate system. Thus, the ε-th time component of $\mathbf{P}$, denoted by $\mathbf{P}_\varepsilon$, can be obtained by averaging the ε-th anti-diagonal components of $\mathbf{C}$ and represented as

$$\mathbf{P}_\varepsilon = \frac{1}{|\Omega_\varepsilon(\mathbf{C})|} \sum_{(u,v)\in\Omega_\varepsilon(\mathbf{C})} \mathbf{c}_{u,v}\ , \tag{30}$$

where $\Omega_\varepsilon(\mathbf{C})$ represents the index set of the ε-th anti-diagonal components in $\mathbf{C}$. Additionally, $|\Omega_\varepsilon(\mathbf{C})|$ represents the operator for calculating the number of elements in $\Omega_\varepsilon(\mathbf{C})$.

We summarize our scheme for BVP matrix estimation in Algorithm 4.

E. HR ESTIMATION IN TIME DOMAIN

In our method, the HR is estimated based on the beat-to-beat peak period analysis of the estimated BVP signal.

1) BVP SIGNAL EXTRACTION

First, we extract the BVP signal from the estimated BVP matrix $\mathbf{P}$. As $\mathbf{P}$ is expected to represent the same BVP dynamics in its all row vectors (i.e., throughout the facial patches), the rank-1 approximation of $\mathbf{P}$ encourages further noise removal of the BVP signal to be estimated. Specifically, the refined BVP matrix, denoted by $\mathbf{X}$, can be obtained by solving the following energy minimization problem:

$$\min_{\mathbf{X}}\ \|\mathbf{X}-\mathbf{P}\|_{\mathrm{F}}^2 \quad \text{s.t. } \mathrm{rank}(\mathbf{X}) = 1\ . \tag{31}$$

As the rank of $\mathbf{X}$ becomes 1, the BVP signal $\mathbf{x}$ can be obtained by extracting the arbitrary row component of $\mathbf{X}$.

2) BEAT-TO-BEAT PEAK PERIOD ANALYSIS

We perform peak detection on $\mathbf{x}$ to obtain the peak locations $\{z_h\}_{h=1}^{\varrho}$ (ϱ is the number of detected peaks in $\mathbf{x}$). By using $\{z_h\}_{h=1}^{\varrho}$, we calculate the average inter-beat interval $\mathcal{I}$ as

$$\mathcal{I} = \frac{\sum_{h=2}^{\varrho}(z_h - z_{h-1})}{\varrho - 1}\ . \tag{32}$$

Finally, we obtain the HR $\mathcal{H}$ by converting $\mathcal{I}$ to the beats-per-minute (bpm) unit:

$$\mathcal{H} = 60/\mathcal{I}\ . \tag{33}$$

IV. EXPERIMENT

A. EXPERIMENTAL SETTINGS

1) DATASET

To demonstrate the effectiveness of the proposed method, we conducted experiments using the TokyoTech Remote PPG dataset [20], MR-NIRP dataset [10], and UBFC-rPPG dataset [19], termed "Tokyo," "MR," and "UBFC," respectively. The details of these datasets are summarized in Table 1.

TABLE 1. Details of datasets used in experiments.

	Tokyo [20]	MR [10]	UBFC [19]
# Subjects	9	8	47
# Videos	9	8	50
Resolution	640×480	640×640	640×480
Bit depth	10 bit	10 bit	8 bit
Frame rate	30 fps	30 fps	30 fps
Duration	180 s	180 s	60 s
Illumination	studio	controlled	realistic

We briefly describe all the datasets used in this experiment.

"Tokyo" dataset: This dataset contained RGB videos of nine participants, who were instructed to sit still and perform a handgrip exercise for approximately 1 min in a room. To create this dataset, two studio lights were used; thus, the lighting environment of this dataset was stable. The RGB videos were captured for 3 min at 30 fps with 640 × 480 resolution with 10-bit depth in an uncompressed format. Ground truth BVP signals were recorded by a finger pulse oximeter at 2048 fps.

"MR" dataset: This dataset contained RGB videos of eight participants, who were asked to sit still under controlled illumination. The RGB videos were captured for 3 min using an RGB camera at 30 fps with 640 × 640 resolution and 10-bit depth in an uncompressed format. Ground truth BVP signals were recorded by a finger pulse oximeter at 60 fps.

"UBFC" dataset: This dataset contained 50 RGB videos from 47 participants. The participants sat and played a time-sensitive mathematical game indoors with varying amounts of natural ambient illumination. This dataset was obtained in a more realistic lighting environment than the other datasets because each participant was captured without any specialized lighting systems. Each video was recorded for 1 min using an RGB camera at 30 fps with a 640 × 480 resolution in an uncompressed 8-bit RGB format. Ground truth BVP signals were recorded by a finger pulse oximeter at 30 fps. We utilized 49 RGB videos in this dataset for the experiments because one video file could not be read.

2) COMPARISON METHODS

We compared the proposed method with the following BVP signal extraction methods: DistancePPG [9], SparsePPG [10], SAMC [11], Hierarchical [21], MTTS-CAN [33], and PhysFormer [34]. The DistancePPG [9], SparsePPG [10], and SAMC [11] methods are based on the spatial low-rank approximation approach. Although the SparsePPG method [10] uses a near-infrared video, we used RGB video as the input for [10] to make a fair comparison in this experiment. The Hierarchical method, which was proposed in our previous study [21], is based on a hierarchical estimation manner using the spatial and temporal characteristics of BVP. The MTTS-CAN [33] and PhysFormer [34] methods are

TABLE 2. Quantitative results for mean absolute error (MAE) [bpm], root-mean-square error (RMSE) [bpm], success rate (SR) [%], and Pearson correlation coefficient (PCC). The best and second best scores are represented in bold and, respectively.

Method	MAE (↓)				RMSE (↓)				SR (↑)				PCC (↑)			
	Tokyo	MR	UBFC	Avg.	Tokyo	MR	UBFC	Avg.	Tokyo	MR	UBFC	Avg.	Tokyo	MR	UBFC	Avg.
DistancePPG [9]	7.00	3.71	6.74	6.40	9.89	5.72	11.95	10.91	67.6	82.5	77.3	76.6	0.50	0.58	0.48	0.50
SparsePPG [10]	128.17	32.13	76.92	78.50	137.56	46.02	83.46	137.60	6.6	27.6	4.2	7.4	0.17	0.02	0.13	0.12
SAMC [11]	8.45	4.96	6.57	6.63	14.43	8.86	10.65	10.95	62.1	82.3	71.5	71.5	0.32	0.43	0.45	0.43
MTTS-CAN [33]	9.39	6.67	19.65	16.70	13.00	8.66	23.02	19.90	49.8	70.5	35.1	41.4	0.09	0.49	0.15	0.18
PhysFormer [34]	51.39	16.72	20.54	24.30	57.97	19.51	25.24	29.01	11.0	52.6	46.4	42.3	0.24	0.25	0.07	0.12
Hierarchical [21]	**3.29**	**2.24**	5.20	4.58	**7.31**	5.15	11.32	10.02	**87.9**	**94.1**	86.7	87.8	0.49	**0.70**	0.56	0.57
Ours	4.11	2.36	**3.34**	**3.32**	7.79	**4.17**	**6.46**	**6.37**	83.3	93.9	**90.0**	**89.6**	**0.56**	0.69	**0.66**	**0.65**

deep-learning-based methods that regress the BVP signal from the input RGB video. In each method, the HR was estimated using beat-to-beat peak period analysis in the time domain.

In this comparison, we set the size of the time window N as 5 s. The time window was moved such that it overlapped its neighbors by 4 s. For the deep learning-based methods [33], [34], we used pre-trained models published by these authors[1],[2], for fair comparison. By conducting preliminary experiments, we set the parameters for the proposed method as $\mu_1 = 4$, $\mu_2 = 1$, $\nu_l = 0.7$ Hz, $\nu_h = 4$ Hz, and $D = 100$. We also ensured that the control parameters of the other methods were optimal.

3) EVALUATION METRICS

Similar to prior works [9], [10], [11], [21], [33], [34], we quantitatively evaluated the results using the root-mean-square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (PCC) between the estimated and ground truth HRs. To obtain the ground truth HRs, we resampled the ground truth BVP signal to match the frame rate of the captured RGB video and performed beat-to-beat peak period analysis, similar to the approach described in Section III-E. In addition, we evaluated the success rate (SR) of HR estimation, which is the proportion of the number of successful HR estimation results to the total number of results. Following previous methods [12], [23], the HR estimation was considered successful if the HR estimation error was below a certain threshold (±5 bpm). Furthermore, we assessed the HR estimation performance using Bland–Altman analysis [40], [41], a data-plotting method for evaluating the agreement between the estimated and ground truth HRs, where the plots in which the measurements are narrowly distributed around zero exhibit better performance.

B. RESULTS

1) RESULTS FOR HR ESTIMATION

Table 2 shows the quantitative results for the MAE, RMSE, SR, and PCC for all datasets. It can be seen that our method exhibited higher accuracy than the comparison methods. Fig. 5 presents the Bland–Altman plots of these methods. The proposed method showed better accuracy than the comparison methods because the plots for the proposed method are narrowly distributed around zero.

Herein, we discuss the results for each dataset. For the Tokyo and MR datasets, which were constructed in stable illumination environments, the other methods [9], [11], [21], especially the hierarchical method [21], were comparable to our method. We conjecture that in a stable illumination scene, the BVP propagation could be clearly observed in every facial patch at the same time. Therefore, the spatial low-rank approximation [9], [11] or hierarchical estimation [21] worked well, leading to accurate HR estimation. Conversely, the HR estimation performance realized using the deep learning-based methods [33], [34] was less accurate than that obtained using the above-mentioned methods. We reason that these methods regressed the BVP signals using their trained model; if the input sequences differ from those used to train their models, the BVP regression performance was significantly degraded due to the overfitting to the training dataset.

In the UBFC dataset, which was acquired in a realistic scene with varying illumination components, the performance of the spatial low-rank approximation methods and hierarchical method [9], [11], [21] was less accurate, primarily because they failed to distinguish between noise and the BVP component. Conversely, the proposed method outperformed the other methods, indicating that our spatio-temporal analysis based on physics-informed DMD contributed to improving the HR estimation performance.

2) COMPARISON IN TIME-SERIES HR ESTIMATION RESULTS

Examples of the time-series HR estimation results are shown in Fig. 6. The results for the stable illumination scene (Fig. 6 (a) and (b)) demonstrated that both our method and the compared methods achieved accurate performance (third row) because patch signals were temporally stable without fluctuating noise components (second row). On the other hand, when an illumination fluctuation occurs, as in the sequences depicted in Fig. 6 (c), the compared methods produced inconsistent and less accurate HR estimation outcomes. In contrast, the proposed method achieved consistent and accurate HR estimation. These results indicate that our

[1] https://github.com/ZitongYu/PhysFormer

[2] https://github.com/xliucs/MTTS-CAN

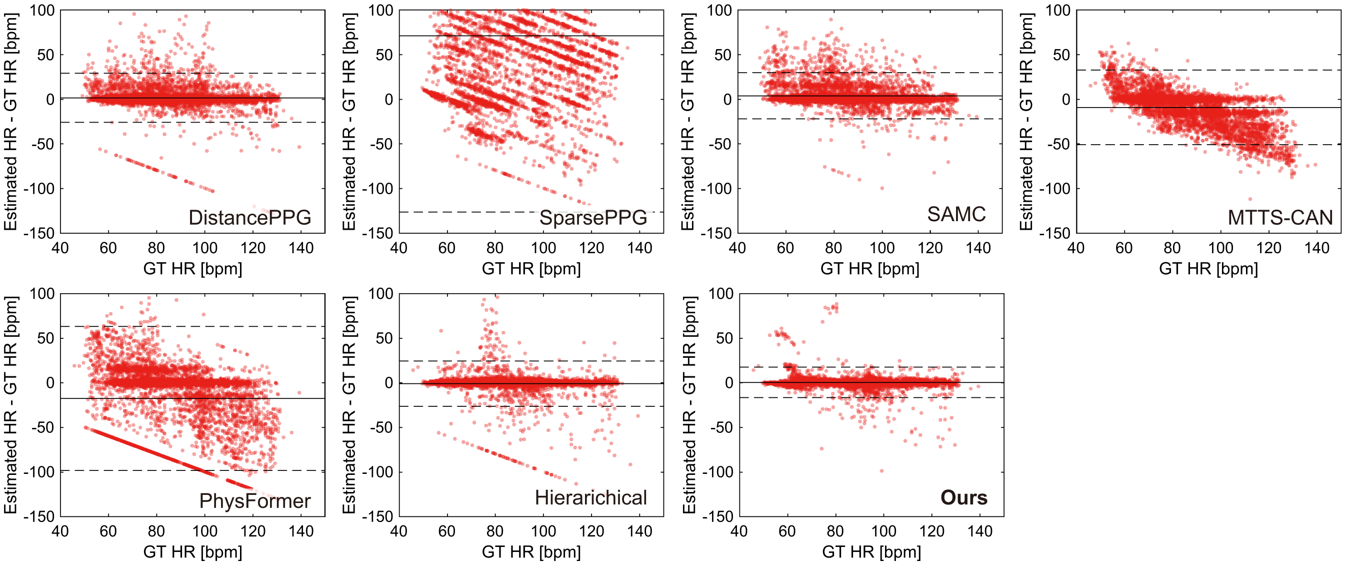
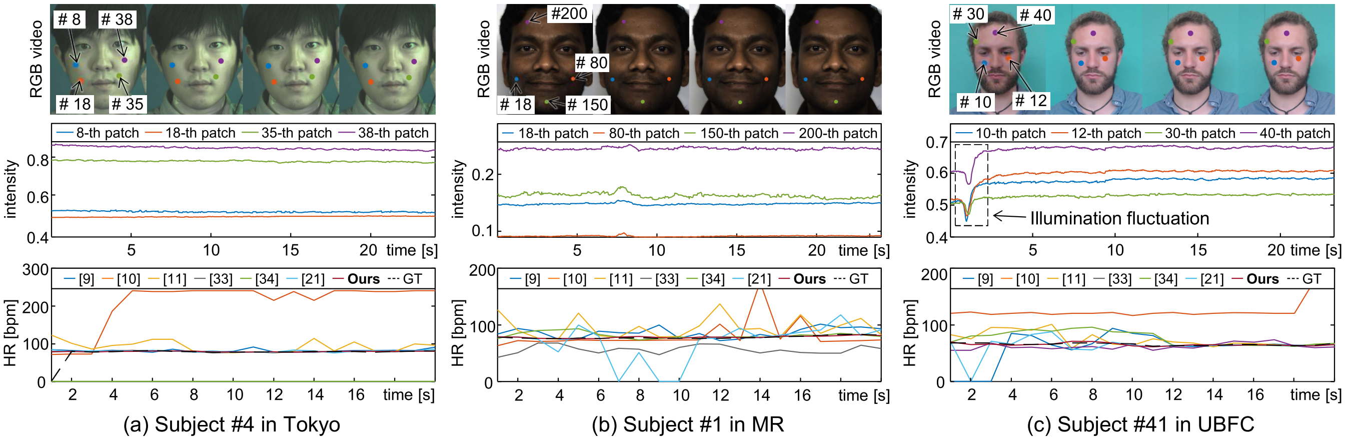


FIGURE 5. Quantitative comparisons using bland–altman plots for all of the participants in all the datasets. In each figure, the solid line shows the mean error and the dashed line indicates the 95 % limits of agreement between the estimated and ground truth HRs.



(a) Subject #4 in Tokyo

(b) Subject #1 in MR

(c) Subject #41 in UBFC

FIGURE 6. Comparison results for time-series variations in estimated HR: The top row depicts RGB video sequences with selected patches (shown as colored circles); the middle row represents time-series signals extracted from selected patches, as indicated by the circles with the corresponding colors in the top row; and the bottom row represents the time-series variations in the estimated HRs ("GT" indicates ground truth HR).

unified spatio-temporal analysis scheme operated well for estimating the BVP signal and HR even in varying illumination scenes.

C. ANALYSIS

1) SIGNIFICANCE AND COMPLEXITY OF EACH MODULE

We investigated the significance and complexity of each module employed in the proposed framework. We list the details of the methods used for this analysis as follows.

(i) Naïve DMD: First, we directly applied naïve DMD to the chrominance matrix $\mathbf{Y}$ and selected the most-dominant DMD mode as the BVP component, i.e., $\mathbf{P}$. Then, we estimated the HR by beat-to-beat peak period analysis as described in Section III-E. We refer to this method as *"Baseline."*

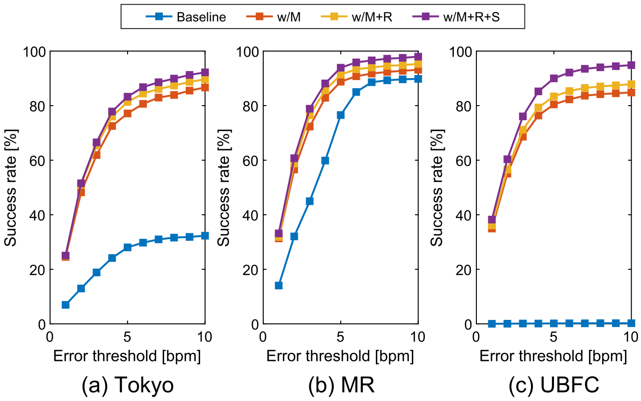
(ii) Physics-informed DMD in Time-Delay Coordinate System: Second, we applied physics-informed DMD to the observation matrix $\mathbf{H}$ in the time-delay coordinate system (Section III-B) to investigate the effects of modeling BVP dynamics exhibiting nonlinearity and quasi-periodicity in the time domain. The HR was estimated using beat-to-beat peak period analysis, the same approach as that used in the "Baseline" method. We refer to this method as *"w/ M."*

(iii) (ii) + Mode Refinement: Third, we added the DMD mode refinement scheme (Section III-C) to "w/ M" to investigate the effect of DMD mode refinement on BVP signal extraction. We refer to this method as *"w/ M+R."*

(iv) (iii) + Adaptive Mode Selection: Finally, we added the adaptive mode selection scheme (Section III-D1) to "w/ M+R" to investigate the effect of incorporating knowledge of the temporal frequency range of BVP. We refer to this

TABLE 3. Impact of each module employed in our framework. We evaluated the performance using the MAE [bpm]. The best scores are represented in bold.

Method	Tokyo	MR	UBFC	Avg.
Baseline	81.52	8.89	96.73	84.01
w/ M	7.75	5.29	13.06	11.39
w/ M+R	6.07	4.13	9.96	8.72
w/ M+R+S	**4.11**	**2.36**	**3.34**	**3.32**



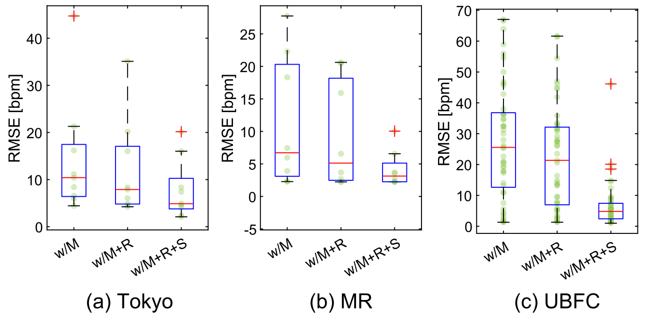
(a) Tokyo (b) MR (c) UBFC

FIGURE 7. Impact of each module employed in our framework in each dataset. We evaluated the performance using the SR curve.

method as *"w/M+R+S."* Note that this approach is equivalent to the proposed full framework.

A) Significance: We investigated the impact of each module on the HR estimation performance. Table 3 presents the comparison results obtained using the MAE metric. It can be seen that the estimation performance was improved as more modules were equipped.

Fig. 7 presents the SR curve, an SR plot obtained by varying the SR threshold from ±1 to ±10 bpm for each method. The horizontal and vertical axes represent the SR threshold and corresponding SR obtained, respectively. For each dataset, each scheme contributed to improving the HR estimation performance. In particular, in the UBFC dataset, the comparison of "Baseline" and "w/M" indicates that our BVP dynamics modeling significantly contributed to performance improvement. In the UBFC dataset, which contains many scenes with varying illumination components, the "Baseline" approach was found to have difficulty estimating the HR accurately, regardless of how the SR threshold was set. According to these results, our modeling of BVP dynamics exhibiting nonlinearity and quasi-periodicity contributed significantly to HR estimation performance.

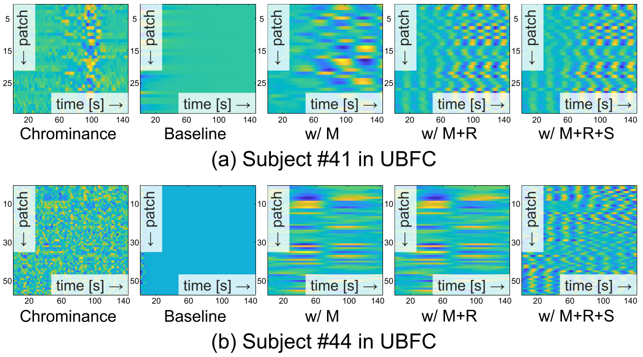


(a) Tokyo (b) MR (c) UBFC

FIGURE 8. Impact analysis of each module employed in our framework in each dataset. We evaluated the performance using box plot analysis. Note that each vertical axis in each figure has a different scale.

TABLE 4. Impact analysis of each module using median (denoted as Med.) and IQR of RMSE values obtained from box plot analysis. The best scores are represented in bold.

Method	Tokyo		MR		UBFC	
	Med.	IQR	Med.	IQR	Med.	IQR
Baseline	130.53	53.43	7.65	25.14	99.30	22.07
w/ M	10.41	11.05	6.71	17.20	25.56	24.16
w/ M+R	7.89	12.25	5.13	15.69	21.36	25.18
w/ M+R+S	**4.88**	**6.49**	**3.14**	**2.87**	**4.80**	**5.04**



(a) Subject #41 in UBFC

(b) Subject #44 in UBFC

FIGURE 9. Examples of estimated BVP matrices.

Fig. 8 presents comparison results obtained using box plot analysis of the RMSE values of each subject in each dataset. Table 4 shows the median and interquartile range (IQR) of the RMSE values obtained with each method. The red horizontal bar in each box plot represents the median of RMSE values. The bottom and top ends of each box, denoted as Q_1 and Q_3, indicate the 25th and 75th percentiles of the RMSE values, respectively. Each vertical dot line (whisker) reaches from the minimum to the maximum RMSE values in the range from $Q_1 + 1.5(Q_3 - Q_1)$ to $Q_3 + 1.5(Q_3 - Q_1)$. Note that $Q_3 - Q_1$ means IQR. The red plus signs represent outliers outside the boundaries of the whiskers. The green circle markers represent the individual data points (i.e., RMSE values). A narrower IQR and lower median of RMSE indicate better precision and accuracy, respectively. Note that the results for "Baseline" are not shown because the RMSEs obtained with this method are considerably higher than the others, making it difficult to plot them on the same scale.

Fig. 8 and Table 4 demonstrated that the median of RMSE obtained with our full framework was the lowest among those obtained using the considered methods. Furthermore, the IQR obtained with our full framework was narrower than those resulting from the other methods for all datasets. These results indicate that our BVP dynamics modeling, DMD mode refinement, and adaptive mode selection scheme

TABLE 5. Example of computation time comparison results for subject #41 in UBFC. The difference in computation time from the left-adjacent method is shown in parentheses.

# of trial	Baseline	w/ M	w/ M+R	w/ M+R+S
#1	0.28 s	0.30 s (+0.02 s)	175.92 s (+175.62 s)	175.69 s (-0.22 s)
#2	0.13 s	0.15 s (+0.02 s)	178.38 s (+178.23 s)	179.95 s (+1.57 s)
#3	0.05 s	0.09 s (+0.04 s)	175.31 s (+175.22 s)	179.16 s (+3.85 s)
#4	0.03 s	0.07 s (+0.04 s)	177.45 s (+177.38 s)	178.95 s (+1.50 s)
#5	0.06 s	0.10 s (+0.04 s)	178.02 s (+177.92 s)	178.85 s (+0.83 s)
Avg.	0.11 s	0.14 s (+0.03 s)	177.02 s (+176.87 s)	178.52 s (+1.51 s)

contribute to improving the precision and accuracy of HR estimation.

Fig. 9 provides examples of the input chrominance matrix **Y** and estimated BVP matrix **P** for visual comparison. The quasi-periodic temporal characteristics of the BVP can be observed in **P** more clearly, as more modules are equipped.

Consequently, each module employed in the proposed framework contributed to the performance enhancement of BVP signal extraction, leading to accurate HR estimation.

B) Complexity: We investigated the computational complexity of our method by measuring the computation time required to execute each module employed in the proposed framework. In fact, because our method solves an alternating iterative optimization problem, rigorously evaluating the computational complexity by analyzing the number of elementary operations using big-o notation is difficult. Therefore, we evaluated the complexity of our method by measuring the computation time to process each method. This analysis was run on MATLAB R2021a installed on a Windows PC with an Intel Core i9-10980XE 3.00 GHz and 64 GB RAM. Considering variations in the computation time, we ran this analysis five times and calculated the average computation time.

Table 5 shows an example of the computation time comparison results. It can be seen that the computation time increased as the number of equipped modules increased. In particular, the difference in computation time between "w/M" and "w/M+R" indicates that the DMD mode refinement scheme was the most computationally intensive among all the modules. In this scheme, as described in Section III-C, we solve alternating iterative optimization problems (Eqs. (20) and (21)) by using a linear solver (i.e., PDIP method), which operates with polynomial complexity [42]. However, because our DMD mode refinement scheme is performed in the high-dimensional space (i.e., time-delay coordinate system), the PDIP method would require expensive computations. This is primarily because the PDIP is a second-order optimization technique involving large Hessian matrix multiplications.

In the future, we plan to adopt a first-order optimization technique that can solve the problem with less computational complexity, such as an iterative shrinkage/thresholding algorithm or the alternating direction method of multipliers. By reducing the computational complexity of estimating HR, our method is expected to be applicable to real scenarios such as health monitoring.

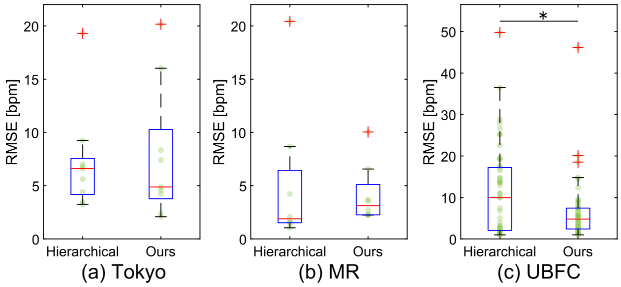


FIGURE 10. Comparison with the hierarchical method in each dataset using box plot analysis. The symbol * indicates a significant difference (i.e., $p < 0.05$) based on the t-test results. Note that each vertical axis in each figure has a different scale.

TABLE 6. Comparison with the hierarchical method using median (denoted as Med.) and IQR of the RMSE values obtained from box plot analysis. The symbol * indicates a significant difference ($p < 0.05$) based on the t-test results. n denotes the number of videos in each dataset (i.e., sample size).

	Tokyo (n = 9)			MR (n = 8)			UBFC (n = 50)		
	Med.	IQR	p-value	Med.	IQR	p-value	Med.	IQR	p-value
[21]	6.61	**3.39**	0.86	**1.89**	4.93	0.71	9.97	15.17	0.01*
Ours	**4.88**	6.49		3.14	**2.87**		**4.80**	**5.04**	

2) DETAILED COMPARISON WITH OUR EARLIER STUDY "HIERARCHICAL"

Here, we analyze the HR estimation performance of our proposed method and our earlier method "Hierarchical" [21] in greater detail. As shown in Section IV-B, our proposed method performed accurate HR estimation. However, in some datasets, the proposed method and "Hierarchical" produced comparable results. To clarify the reason for this finding, we conducted a more detailed analysis. Specifically, we assessed the accuracy and precision of the "Hierarchical" and our method using a box plot analysis of the RMSEs of each subject in each dataset. Furthermore, we conducted Welch's two-tailed t-test (significance level: $p < 0.05$) on the RMSE values of "Hierarchical" and our method to assess whether there is a significant difference between them.

Fig. 10 shows the comparison results using a box plot of RMSE values for each subject in each dataset. Table 6 lists the median and interquartile range (IQR) of RMSE and p-value. In the Tokyo and MR datasets, we believe that the performance of our method was comparable to that of "Hierarchical" based on the comparison of the medians and IQRs obtained with each method. According to the t-test results, the difference between "Hierarchical" and our method is insignificant ($p = 0.01$).

We discuss why "Hierarchical" performed comparably to our method in the Tokyo and the MR datasets. In practice, the Tokyo and MR datasets were constructed in a stable illumination scene, where the temporal characteristics of the BVP could be clearly observed from each facial patch. In such scenes with less noise component, ‘‘Hierarchical’’ can work accurately to model the temporal characteristics of the BVP. Therefore, in the stable illumination scene, our method and ‘‘Hierarchical’’ showed comparable HR estimation performance.

Conversely, for the UBFC dataset, the median and IQR obtained with our method are lower and narrower than those of ‘‘Hierarchical,’’ indicating that our method achieved superior performance. Furthermore, the t-test results revealed a significant difference between the RMSEs obtained from ‘‘Hierarchical’’ and our method. Based on the above results, we can contend that our method produced quantitatively better results than ‘‘Hierarchical.’’ In the UBFC dataset that includes varying illumination scenes, extracting a reliable BVP signal using the ‘‘Hierarchical’’ is difficult. This is because in the first stage of ‘‘Hierarchical,’’ time-series modeling of the BVP is performed, whereas the spatial similarity of the BVP is discarded. Therefore, the estimation accuracy of the BVP candidates in the first stage is degraded, leading to performance degradation in the overall framework of ‘‘Hierarchical.’’ Furthermore, in varying illumination scenes, the autoregressive time-series modeling utilized in ‘‘Hierarchical’’ is insufficient to extract reliable BVP signals. We reason that varying illumination components can also be modeled by an autoregressive model. Thus, BVP component and noise due to varying illuminations cannot be distinguished by ‘‘Hierarchical.’’ By contrast, our method is a spatio-temporal analysis approach incorporating quasi-periodic and nonlinear dynamics of the BVP based on physics-informed DMD framework. This framework enables the unified spatio-temporal modeling of the BVP and thus enables distinction between the BVP and noise components. Therefore, our method showed better HR estimation performance than ‘‘Hierarchical’’ in the UBFC dataset.

V. CONCLUSION

A. SUMMARY

We proposed a BVP signal extraction method for HR estimation that incorporates medical knowledge of BVP dynamics. Based on the DMD framework, we exploited the BVP characteristics in the spatial and temporal domains in a unified manner for HR estimation. To analyze the BVP dynamics exhibiting nonlinear and quasi-periodic properties, we performed physics-informed DMD on the time-series signals in the time-delay coordinate system. The estimated spatio-temporal structures attributable to the BVP signal were refined using an optimization framework regularized with the spatial similarity of the BVP. The BVP signal and HR were estimated by inverse time-delay embedding of the outcome obtained using our adaptive best DMD mode selection based on the medical knowledge of the HR frequency range. Through experiments using public datasets, we demonstrated the effectiveness of the proposed method.

B. FUTURE WORK

In the proposed approach, the BVP signal is estimated based on its quasi-periodicity in the time domain and spatial similarity over the face region. If quasi-periodic illumination variations, similar to those of the BVP, are subjected to the face, our method may fail in HR estimation because this phenomenon deviates from the assumption made in this study. To address this remaining issue, we will investigate methods of removing such noise component by incorporating multispectral (such as ultraviolet and near-infrared) information, as in [43], [44], [45], and [46].

REFERENCES

[1] D. McDuff, S. Gontarek, and R. Picard, ‘‘Remote measurement of cognitive stress via heart rate variability,’’ in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2014, pp. 2957–2960.

[2] M. Burzo, D. McDuff, R. Mihalcea, L.-P. Morency, A. Narvaez, and V. Perez-Rosas, ‘‘Towards sensing the influence of visual narratives on human affect,’’ in *Proc. ACM Int. Conf. Multimodal Interact.*, Oct. 2012, pp. 153–160.

[3] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, ‘‘Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics,’’ *Sci. Rep.*, vol. 4, no. 1, pp. 1–10, May 2014.

[4] *Nevermind*. Accessed: Nov. 10, 2022. [Online]. Available: https://nevermindgame.com/

[5] J. Allen, ‘‘Photoplethysmography and its application in clinical physiological measurement,’’ *Physiolog. Meas.*, vol. 28, no. 3, pp. R1–R39, Mar. 2007.

[6] Y. Sun and N. Thakor, ‘‘Photoplethysmography revisited: From contact to noncontact, from point to imaging,’’ *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 463–477, Mar. 2016.

[7] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, ‘‘Remote plethysmographic imaging using ambient light,’’ *Opt. Exp.*, vol. 16, no. 26, pp. 21434–21445, 2008.

[8] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, ‘‘Video-based heart rate measurement: Recent advances and future prospects,’’ *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.

[9] M. Kumar, A. Veeraraghavan, and A. Sabharwal, ‘‘DistancePPG: Robust non-contact vital signs monitoring using a camera,’’ *Biomed. Opt. Exp.*, vol. 6, no. 5, pp. 1565–1588, Apr. 2015.

[10] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1272–1281.

[11] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2396–2404.

[12] A. Lam and Y. Kuno, ‘‘Robust heart rate measurement from video using select random patches,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3640–3648.

[13] W. B. Murray and P. A. Foster, ‘‘The peripheral pulse wave: Information overlooked,’’ *J. Clin. Monitor.*, vol. 12, no. 5, pp. 365–377, Sep. 1996.

[14] M. Elgendi, ‘‘On the analysis of fingertip photoplethysmogram signals,’’ *Current Cardiol. Rev.*, vol. 8, no. 1, pp. 14–25, Jun. 2012.

[15] S. Pan and K. Duraisamy, ‘‘On the structure of time-delay embedding in linear models of non-linear dynamical systems,’’ *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 30, no. 7, pp. 073135-1–073135-29, 2020.

[16] K. P. Champion, S. L. Brunton, and J. N. Kutz, ‘‘Discovery of nonlinear multiscale systems: Sampling strategies and embeddings,’’ *SIAM J. Appl. Dyn. Syst.*, vol. 18, no. 1, pp. 312–333, Jan. 2019.

[17] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, ‘‘Chaos as an intermittently forced linear system,’’ *Nature Commun.*, vol. 8, no. 1, pp. 1–15, May 2017.

[18] N. Sviridova and K. Sakai, ‘‘Human photoplethysmogram: New insight into chaotic characteristics,’’ *Chaos, Solitons Fractals*, vol. 77, pp. 53–63, Aug. 2015.

[19] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019.

[20] 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. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6525–6528.

[21] K. Kurihara, Y. Maeda, D. Sugimura, and T. Hamamoto, "Blood volume pulse signal extraction based on spatio-temporal low-rank approximation for heart rate estimation," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.

[22] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 4264–4271.

[23] R. Spetlík, V. Franc, J. Cech, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 3–6.

[24] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.

[25] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[26] L. Wei, Y. Tian, Y. Wang, T. Ebrahimi, and T. Huang, "Automatic webcam-based human heart rate measurements using Laplacian eigenmap," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2012, pp. 281–292.

[27] X. Niu, H. Han, S. Shan, and X. Chen, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Dec. 2018, pp. 562–576.

[28] K. Kurihara, D. Sugimura, and T. Hamamoto, "Non-contact heart rate estimation via adaptive RGB/NIR signal fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 6528–6543, 2021.

[29] J. Cheng, X. Chen, L. Xu, and Z. J. Wang, "Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1422–1433, Sep. 2017.

[30] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[31] X. Niu, S. Shan, H. Han, and X. Chen, "RhythmNet: End-to-end heart rate estimation from face via spatial–temporal representation," *IEEE Trans. Image Process.*, vol. 29, pp. 2409–2423, 2020.

[32] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 349–365.

[33] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 19400–19411.

[34] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4176–4186.

[35] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1373–1384, May 2021.

[36] E. Sánchez-Lozano, G. Tzimiropoulos, B. Martinez, F. De la Torre, and M. Valstar, "A functional regression approach to facial landmark tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2037–2050, Sep. 2018.

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

[38] P. J. Baddoo, B. Herrmann, B. J. McKeon, J. N. Kutz, and S. L. Brunton, "Physics-informed dynamic mode decomposition," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 479, no. 2271, Mar. 2023.

[39] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz, "Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition," *J. Neurosci. Methods*, vol. 258, pp. 1–15, Jan. 2016.

[40] 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, Feb. 1986.

[41] J. S. Krouwer, "Why Bland–Altman plots should use X, not $(Y + X)/2$ when X is a reference method," *Statist. Med.*, vol. 27, no. 5, pp. 778–780, 2008.

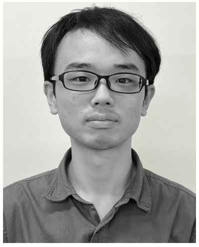
[42] L. G. Khachiyan, "Polynomial algorithms in linear programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 1, pp. 53–72, Jan. 1980.

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

[44] W. Wang, A. C. den Brinker, and G. de Haan, "Discriminative signatures for remote-PPG," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 5, pp. 1462–1473, May 2020.

[45] 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. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4534–4538.

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



KOSUKE KURIHARA (Graduate Student Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from the Tokyo University of Science, Japan, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree. His research interests include image processing and biosignal processing.

YOSHIHIRO MAEDA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information engineering from the Nagoya Institute of Technology, Japan, in 2013, 2015, and 2019, respectively. He became an Assistant Professor with the Tokyo University of Science, Japan, in 2019. His research interests include image processing and multispectral sensing.

DAISUKE SUGIMURA (Member, IEEE) 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. His research interests include computer vision and computational imaging.

TAKAYUKI HAMAMOTO (Member, IEEE) received the B.E. and M.E. 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, in 1997. He is currently a Professor with the Department of Electrical Engineering, Tokyo University of Science. His research interests include image processing, computer vision, and computational image sensors.