Blood Volume Pulse Signal Extraction based on Spatio-Temporal Low-Rank Approximation for Heart Rate Estimation

Kosuke Kurihara[1], Yoshihiro Maeda[1], Daisuke Sugimura[2], and Takayuki Hamamoto[1]
[1]*Tokyo University of Science, Tokyo, 125-8585, Japan* [2]*Tsuda University, Tokyo, 187-8577, Japan*
[1]4321701@ed.tus.ac.jp

***Abstract*—We propose a novel blood volume pulse (BVP) signal extraction method for heart rate estimation that incorporates the self-similarity properties of BVP in the spatial and temporal domains. The main novelty of the proposed method is the incorporation of the temporal self-similarity of BVP via low-rank approximation in the time-delay coordinate system for BVP signal extraction. To make a low-rank approximation of BVP in the time domain, we introduce knowledge of linear time-invariant systems, i.e., the autoregressive (AR) model lies in the low-rank subspace in the time-delay coordinate system. In the medical field, it is widely known that BVP has quasi-periodic temporal characteristics owing to the cardiac pulse and exhibits self-similarity properties in the temporal domain. Hence, we model the temporal behavior of BVP as an AR process, allowing for a low-rank approximation of BVP in the time-delay coordinate system. Low-rank approximation of BVP in the time and spatial domains enables reliable BVP signal extraction, resulting in accurate heart rate estimation. The experiments demonstrate the effectiveness of the proposed method.**

***Index terms*—** non-contact vital sensing, heart rate estimation, time-delay embedding, low-rank approximation

I. INTRODUCTION

Heart rate (HR) provides insights into the person's physiological and emotional state. HR can be defined as the average speed of the heartbeat that can be measured by counting the number of the cardiac pulses appearing in a certain time window.

Traditional methods for HR estimation require contact-type sensors, such as electrocardiograms and pulse oximetry sensors. These methods are widely used for their good accuracy; however, the restrictions associated with these contact-type sensors make subjects uncomfortable. Therefore, it is desirable to develop non-contact HR estimation.

This article has been accepted for publication in IEEE International Conference on Visual Communications and Image Processing (VCIP) 2022. 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/VCIP56404.2022.10008871

978-1-6654-7592-1/22/$31.00 ©2022 IEEE

In the last decade, numerous methods for non-contact HR estimation using cameras have been proposed. In fact, the blood volume pulse (BVP) associated with the cardiac pulse causes temporal and periodic changes in skin color in facial videos. Hence, BVP can be measured by analyzing such temporal changes in a facial video. Once the BVP signal has been measured, the HR outcome can also be obtained.

Previous methods extracted the BVP signal for HR estimation based on a frequency analysis [1]–[3]. In these methods, it is assumed that the BVP contains the dominant frequency component derived from the cardiac pulse. Based on this assumption, HR is estimated by searching for the maximum frequency component of the extracted BVP signal. However, these frequency-domain methods have the inherent limitation that the accuracy of HR estimation is limited by the frequency resolution of the BVP signal. According to the previous literature [1]–[3], frequency-domain methods require a long observation time (10 - 30 s) to obtain the fine frequency resolution required for accurate HR estimation. This limitation reduces HR estimation performance in cases where HR is prone to sudden changes (e.g., instantaneous emotions, sports, and exercises). Hence, it is difficult to apply frequency-domain methods to a wide variety of scenarios.

Unlike frequency-domain methods, time-domain methods allow accurate HR estimation, even when HR changes rapidly [4]. This advantage is mainly because the frequency resolution remains the same, no matter how short the observation time. However, HR estimation accuracy is significantly degraded due to noise arising from local facial movements (e.g, facial expression) and non-uniform fluctuations in illumination [5]. In fact, the temporal variation in skin color arising from the BVP is quite subtle; the pixel intensity attributable to the BVP is merely varied in less than 2 bits of the analog-to-digital converter of the camera [5].

To overcome the aforementioned problems, the authors of [6] proposed a time-domain method using the spatial similarity of the BVP signal based on multiple facial patches representation. This method [6] assumed that the BVP signal could be observed similarly in neighboring patches because the blood flows over the facial region with approximately the same timing. Based on this assumption, they claimed that the BVP signal could be represented by a low-rank approximation in the spatial domain. Specifically, the robust principal component

analysis was performed on a set of facial patch signals to obtain low-rank components, including a BVP signal with less noise.

However, when similar noise is imposed on all facial patches due to motion of subjects or fluctuations in illumination, the BVP signal is difficult to extract accurately using this method [6]. This is primarily because many of the imposed noise components appear in the low-rank component, which is expected to contain a less noisy BVP signal. Hence, the HR estimation performance also declines.

In this study, we propose a novel BVP signal extraction method that incorporates the spatio-temporal characteristics of the BVP signal to accurately estimate HR in a non-contact manner. In the medical field, BVP has quasi-periodic characteristics owing to the cardiac pulse [7], indicating that it contains self-similarity properties in the temporal domain and thus can be represented as an autoregressive (AR) process. The main novelty of the proposed method is the incorporation of the temporal self-similarity of the BVP signal via a low-rank approximation in the time-delay coordinate system. In the realization theory of linear time-invariant system, it is well known that the order of the AR process is related to the rank of a time-series signal represented in the time-delay coordinate system. This suggests that a low-rank approximation of the BVP signal in the time-delay coordinate system can be achieved. By performing a low-rank approximation of the latent BVP signal in the spatial and temporal domains, the proposed method enables accurate extraction of the BVP signal, leading to accurate HR estimation, even with a short observation time.

II. Related Work

The framework for HR estimation from videos is mainly composed of BVP signal extraction and HR estimation from the extracted BVP signal. Once the accurate BVP signal has been extracted, the accurate HR outcome can be obtained. Thus, many researchers proposed BVP signal extraction scheme for accurate HR estimation.

There are methods for BVP signal extraction based on frequency analysis [1], [5]. Kumar *et al.* proposed a BVP extraction scheme using weighted average of multiple patch observations [5]. In the method [5], the weights for fusing multiple patch observations were determined based on the assumption that the dominant frequency component would be obtained if the patch signal contained components attributable to a cardiac pulse. Nowara *et al.* [1] proposed a noise reduction scheme in the frequency domain using multiple patch observations for HR estimation. They assumed that the BVP candidates obtained from multiple patches would contain dominant frequency components derived from the cardiac pulse. Based on this assumption, the joint sparsity of the frequency components among multiple spatial patches was used to denoise the BVP signal. Because the aforementioned methods rely on analysis in the frequency domain, HR estimation performance is limited by the frequency resolution

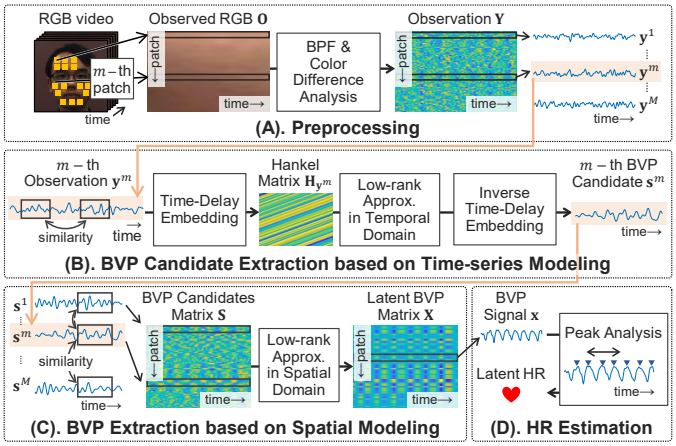

Fig. 1: Overview of the proposed method.

of patch signals. In practice, these methods require long-term observations (10 - 30 s) to obtain the fine frequency resolution for accurate HR estimation. This limitation degrades HR estimation performance when the HR changes suddenly, such as instantaneous changes in emotion, exercise, and sports [8].

Several studies have been reported on BVP signal extraction in the time domain [6]. Tulyakov *et al.* [6] introduced a matrix completion scheme in the time domain for BVP signal extraction, based on a spatial low-rank approximation between multiple patches.

The aforementioned time-domain method can alleviate the problems of frequency-domain approaches. However, if all spatial patches are subjected to similar noise, it is difficult to distinguish the noise component from the BVP because the noise component also satisfies the same spatial low rankness as the BVP. In contrast, the proposed method allows for low-rank approximation of the BVP signal in the spatial and temporal domains, alleviating the aforementioned problem of [6].

III. Proposed Method

Figure 1 presents an overview of the proposed method. We first detect the facial region of a subject from an input RGB facial video, and divide it into M facial patches using the previous method [9]. By tracking each facial patch in an input video using [9], we obtain M time-series RGB signals attributable to the latent BVP signal. Note that each patch signal is obtained by averaging the pixel values within the patch.

In the proposed method, we represent a set of observed RGB signals in the matrix form $\mathbf{O} \in \mathbb{R}^{M \times T}$ (T denotes the size of the time window). As described earlier, $\mathbf{O}$ includes the BVP signals as well as the noise caused by changes in the illumination or head movements of the subject. Thus, our objective is to extract the BVP signal accurately from noisy observation $\mathbf{O}$. HR can then be estimated by measuring the beat-to-beat peak positions of the estimated BVP signal. We detail the proposed method in subsequent sections.

A. Preprocessing

We describe the preprocessing procedure for the observed RGB signals, which is used to extract reliable components attributable to the BVP signal, similar to the previous studies [2], [9]. First, we use a band-pass filter to remove the frequency component outside the range of HR. We determine the bandwidth of the band-pass filter as 0.5 - 8 Hz based on the knowledge of normal HR range of a human [10]. We then perform color-difference space conversion for the filtered signals. According to the past study [11], analysis in the color-difference space was effective for BVP extraction from RGB videos. Based on this findings, we project the RGB patch signals into the color-difference space using the method [11]. The processed observation, defined as $\mathbf{Y} \in \mathbb{R}^{M \times T}$, is used in the proposed method and represented as

$$\mathbf{Y} = \mathcal{C}\left[\mathrm{BPF}\left(\mathbf{O}\right)\right] \ , \tag{1}$$

where $\mathcal{C}[\cdot]$ and $\mathrm{BPF}(\cdot)$ denote the operators of color-difference space conversion and band-pass filtering, respectively.

As mentioned earlier, the observation $\mathbf{Y}$ includes the components attributable to the BVP signal as well as the residual noise components; thus, it can be modeled as

$$\mathbf{Y} = \mathbf{S} + \mathbf{E} \ , \tag{2}$$

where $\mathbf{S}$ and $\mathbf{E}$ denote the BVP candidate and residual noise matrices, respectively.

B. BVP Candidate Extraction

In this section, we detail the scheme for BVP candidate extraction based on the temporal characteristics of the BVP signal.

1) Autoregressive Model: Since BVP has quasi-periodic temporal characteristics induced by the cardiac cycle, it can be modeled as a stationary process. According to the past study [12], an AR model is suitable for representing the temporal behaviors of BVP signal. Based on this finding, the AR model is used for the proposed method.

Let the m-th facial patch components of $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{E}$ (i.e, m-th row component of each matrix) be denoted as $\mathbf{y}^m = (y_1^m, y_2^m, ..., y_T^m)$, $\mathbf{s}^m = (s_1^m, s_2^m, ..., s_T^m)$, and $\mathbf{e}^m = (e_1^m, e_2^m, ..., e_T^m)$, respectively. Using y_t^m, s_t^m, and e_t^m, the AR model is formulated as

$$y_t^m = s_t^m + e_t^m, \tag{3}$$

$$s_t^m = \sum_{i=1}^{p} \varphi_i s_{t-i}^m, \tag{4}$$

$$e_t^m = \varepsilon_t \ , \tag{5}$$

where φ_i and p denote the i-th coefficient and order of the AR model, respectively. In addition, e_t^m is represented as white noise ε_t according to [12].

2) Time-Delay Coordinate System: In the realization theory of linear time-invariant system, it is well known that the order of the AR model p is related to the rank of a time-series signal modeled by the AR process in the time-delay coordinate system [13]. This suggests that the time-series signal $\mathbf{s}^m$ and the residual noise components $\mathbf{e}^m$ would be separated using the first p-rank components of the observation $\mathbf{y}^m$ in the time-delay coordinate system.

We first project $\mathbf{y}^m$, $\mathbf{s}^m$, and $\mathbf{e}^m$ into the time-delay coordinate system. In this coordinate system, these are respectively represented as Hankel matrices [14]: $\mathbf{H}_{\mathbf{y}^m}$, $\mathbf{H}_{\mathbf{s}^m}$, and $\mathbf{H}_{\mathbf{e}^m}$. The observations in the time-delay coordinate system, i.e., $\mathbf{H}_{\mathbf{y}^m}$, are modeled as

$$\mathbf{H}_{\mathbf{y}^m} = \mathbf{H}_{\mathbf{s}^m} + \mathbf{H}_{\mathbf{e}^m} \ , \tag{6}$$

where

$$\mathbf{H}_{\mathbf{y}^m} = \begin{bmatrix} y_1^m & y_2^m & \cdots & y_{T-r+1}^m \\ y_2^m & y_3^m & \cdots & y_{T-r+2}^m \\ \vdots & \vdots & \ddots & \vdots \\ y_r^m & y_{r+1}^m & \cdots & y_T^m \end{bmatrix} \ , \tag{7}$$

where r is the dimension of the time-delay embedding.

3) Estimation of BVP Candidates: Based on the aforementioned representation in the time-delay coordinate system, we estimate the latent $\mathbf{H}_{\mathbf{s}^m}$ with the low-rank approximation of $\mathbf{H}_{\mathbf{s}^m}$ as

$$\min_{\mathbf{H}_{\mathbf{s}^m}} \ \|\mathbf{H}_{\mathbf{s}^m} - \mathbf{H}_{\mathbf{y}^m}\|_F \quad \text{s.t. } \mathrm{rank}(\mathbf{H}_{\mathbf{s}^m}) \leqq p, \tag{8}$$

where $\|\cdot\|_F$ and $\mathrm{rank}(\cdot)$ denote the operators that compute the Frobenius norm and the rank of an input matrix, respectively.

To obtain the BVP candidate $\mathbf{s}^m$, we perform inverse time-delay embedding of $\mathbf{H}_{\mathbf{s}^m}$ (i.e, de-Hankelization), which is represented as

$$\mathbf{s}^m = \mathcal{H}^{-1}(\mathbf{H}_{\mathbf{s}^m}), \tag{9}$$

where $\mathcal{H}^{-1}(\cdot)$ denotes the de-Hankelization operator.

We perform the aforementioned processing for each facial patch. The BVP candidate matrix $\mathbf{S}$ can be obtained by stacking each $\mathbf{s}^m$ in a row.

C. BVP Signal Extraction

We estimate the latent BVP matrix $\mathbf{X}$ from the BVP candidate matrix $\mathbf{S}$. According to the past study [6], BVP tends to be similar at neighboring facial patches because the blood flows over the facial region at approximately the same time. Based on this assumption, we estimate $\mathbf{X}$ via the low-rank approximation of $\mathbf{X}$ in the spatial domain. Because BVP would be unique regardless of the facial patches, the rank of $\mathbf{X}$ can be approximated as 1. Hence, we estimate $\mathbf{X}$ by solving the following optimization problem:

$$\min_{\mathbf{X}} \ \|\mathbf{X} - \mathbf{S}\|_F \quad \text{s.t. } \mathrm{rank}(\mathbf{X}) = 1 \ . \tag{10}$$

D. HR Estimation in the Time Domain

To estimate HR, we apply the beat-to-beat peak period analysis on the estimated BVP matrix $\mathbf{X}$. Since rank of $\mathbf{X}$ is 1, the BVP signal, denoted as $\mathbf{x}$, can be obtained from the arbitrary row component of $\mathbf{X}$. We perform the peak detection on $\mathbf{x}$ to obtain the peak locations $\{\tau_k\}_{k=1}^{K}$ (K is the number of detected peaks in $\mathbf{x}$). Using $\{\tau_k\}_{k=1}^{K}$, we calculate the average inter-beat interval d as

$$d = \frac{\sum_{k=2}^{K}(\tau_k - \tau_{k-1})}{K-1} \ . \tag{11}$$

We finally obtain HR by converting d to the beats-per-minute (bpm) unit.

IV. EXPERIMENTS

A. Experimental Settings

1) Dataset: To demonstrate the effectiveness of the proposed method, we conducted experiments using TokyoTech Remote PPG dataset [15], MR-NIRP dataset [1], and UBFC dataset [16], called "Tokyo," "MR," and "UBFC," respectively. The details of these datasets are summarized in Table I.

2) Comparison Methods: We compared the proposed method with the following BVP signal extraction methods: DistancePPG [5], SparsePPG [1], and SAMC [6]. Specifically, DistancePPG [5] and SparsePPG [1] are based on frequency analysis, and SAMC [6] is a time-domain approach. In each method, HR was estimated using beat-to-beat peak period analysis. Note that although SparsePPG method used a near-infrared video [1], we used RGB video as the input for SparsePPG method to make a fair comparison.

We set the size of time window T to 5 s for all methods to evaluate the HR estimation performance over a short-term observation time. The time window was moved such that it overlapped with its neighbors by 4 s. By conducting preliminary experiments, we set the parameters for the proposed method to $p = 6$ and $r = 100$. We also ensured that the control parameters of the other comparison methods were optimal.

3) Evaluation Metrics: We quantitatively evaluated the results using the mean absolute error (MAE). In addition, we evaluated the success rate (SR) of HR estimation by aggregating the outputs for which the difference between the estimated and ground-truth HRs was less than a certain threshold (± 5 bpm). Furthermore, we assessed the HR estimation performance using the Bland-Altman analysis [17], [18], 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.

B. Results

Table II shows the comparison results of the MAE and SR. Each value was obtained by averaging the results of all subjects in each dataset. It can be seen that the proposed method shows the highest accuracy among the comparison methods.

TABLE I: Details of datasets used in experiments.

	Tokyo [15]	MR [1]	UBFC [16]
# Subjects	9	8	47
# Videos	9	8	50
Resolution	640×480	640×640	640×480
Frame rate	30 fps	30 fps	30 fps
Duration	180 s	180 s	60 s

TABLE II: Quantitative results of MAE and SR.

	MAE (bpm)				SR (%)			
	Ours	[6]	[5]	[1]	Ours	[6]	[5]	[1]
Tokyo [15]	**3.7**	6.0	15.3	62.4	**86.4**	71.8	39.8	16.1
MR [1]	**1.9**	2.2	10.3	28.6	**94.2**	92.9	62.9	15.6
UBFC [16]	**4.9**	10.7	9.2	58.3	**75.8**	53.2	62.2	7.4

The results are discussed below: SAMC [6], which is a time-domain approach, showed better results than the other comparison methods [1], [5]. We reason that the time-domain approach was less affected by the frequency-resolution problem, which becomes a crucial issue with frequency-domain methods [1], [5] when the observation time is short. However, the results of SAMC [6] were worse than those of the proposed method. We reason that SAMC used only the similarity of adjacent patch signals to perform low-rank approximation; the noise components added to the facial patches contaminated the low-rank component, resulting in poor HR estimation performance.

The results based on the Bland-Altman plot are shown in Fig. 2. The plots obtained by the proposed method were narrowly distributed around zero, indicating better performance than the other comparison methods.

V. CONCLUSION

We proposed a novel BVP signal extraction method for non-contact HR estimation that incorporates the self-similarity properties of BVP in the spatial and temporal domains. We modeled the temporal behavior of BVP as an AR process, allowing for a low-rank approximation of BVP in the time-delay coordinate system. The low-rank approximation of BVP in the temporal and spatial domains enabled reliable BVP signal extraction, resulting in an accurate heart rate estimation. Through the experiments, we demonstrated the effectiveness of the proposed method.

REFERENCES

[1] 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.

[2] 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.

[3] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. of IEEE ICCV*, 12 2015, pp. 3640–3648.

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

[5] 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.

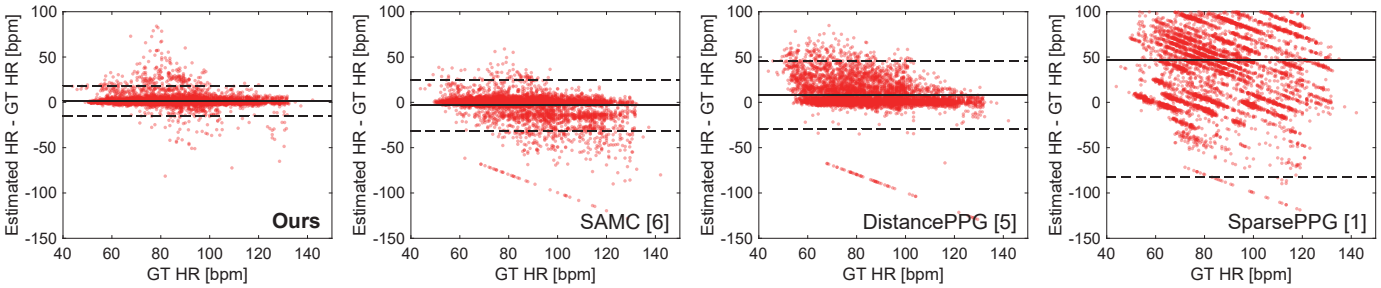

Fig. 2: Results of Bland-Altman plots. The solid and dotted lines represent the mean error and 95% confidential interval, respectively.

[6] 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.

[7] W. B. Murray and P. A. Foster, "The peripheral pulse wave: information overlooked," *Journal of Clinical Monitoring*, vol. 12, pp. 365–377, 1996.

[8] G. Valenza, L. Citi, A. Lanatrk, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics," *Scientific Reports*, vol. 4, 2014.

[9] 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.

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

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

[12] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, pp. 807–831, 2014.

[13] M. Ayazoglu, B. Li, C. Dicle, M. Sznaier, and O. I. Camps, "Dynamic subspace-based coordinated multicamera tracking," in *Proc. of IEEE ICCV*, 2011, pp. 2462–2469.

[14] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proc. of IEEE CVPR*, 2012, pp. 1362–1369.

[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] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, no. 1, pp. 82–90, 2019.

[17] 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.

[18] 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.