Explainable Estimation of Blood Volume Pulse Signals from Video Sequences Using a Combination of Deep Learning Models and Signal Processing Methods

Milena Sobotka(✉), Kamil Kopryk, Muhammad Usman, and Jacek Rumiński

Department of Biomedical Engineering, Gdansk University of Technology, FETI,
80-233 Gdansk, Poland
milena.sobotka@pg.edu.pl
http://www.pg.edu.pl

Abstract. This study evaluates the application of Gradient-weighted Class Activation Mapping (Grad-CAM) to identify key image regions for enhancing pulse wave estimation using traditional signal analysis methods. The approach assumes that Grad-CAM masking enables the selection of relevant image areas, improving the Signal to Noise Ratio (SNR). Experiments were conducted on the PURE dataset using the TS-CAN model and rPPG-Toolbox. Grad-CAM maps identified facial regions most influential for model prediction, allowing the exclusion of areas not contributing to accurate estimation. The study also explored different temporal window sizes and their impact on signal quality. Method evaluation included SNR$_{raw}$, SNR$_{max}$, SNR$_{sum}$, and Hjorth descriptors. Results confirmed that Grad-CAM masking enhances rPPG signal quality, enabling more precise heart rate estimation. Statistical analysis validated the significance of the findings, highlighting the potential of interpretable deep learning methods in signal analysis.

Keywords: blood volume pulse · remote photoplethysmography · signal processing · explainable AI

1 Introduction

Technological advancements have enabled the development of non-invasive methods for monitoring physiological parameters, among which remote photoplethysmography (rPPG) plays a crucial role. This technique utilizes subtle changes in skin color, captured in video recordings, to analyze pulsatile blood flow, allowing for the estimation of heart rate (HR) and blood volume pulse (BVP) [1]. These variations result from small fluctuations in blood volume during the cardiac cycle, which influence the amount of light absorbed by blood vessels and, consequently, cause alterations in skin intensity values. Unlike traditional contact-based methods, such as photoplethysmography (PPG), rPPG eliminates the

P. Ladyzynski et al. (Eds.): NBC/PCBBE 2025, IFMBE Proceedings 131, pp. 75–89, 2025.
https://doi.org/10.1007/978-3-031-96538-8_8

need for attached sensors, significantly enhancing user comfort and enabling continuous, remote monitoring. As a result, it finds wide application in home-based health monitoring systems. Despite its advantages, rPPG still encounters challenges related to physiological factors influencing BVP signal measurement. One of the key obstacles is the spatial variability of the vascular system and the uneven distribution of signal intensity across different regions of interest (ROI) on the face. Furthermore, the low amplitude of these signals makes them particularly vulnerable to motion artifacts and lighting fluctuations, often causing noise levels to exceed the intensity of the useful signal. In response to these challenges, various filtering algorithms and noise reduction techniques have been developed [2–4]. Among the most commonly used methods are those based on independent component analysis (ICA), which allow for the decomposition of RGB signals into independent components, from which those containing relevant heart rate information are selected. Alternatively, models are employed where RGB channels are represented as a linear combination of BVP components and noise, allowing for their separation based on prior assumptions. Deep learning-based methods [5,6] have achieved significant importance in analyzing video data to extract physiological parameters such as HR and BVP. Once trained on large datasets, these models can effectively predict HR components, thereby improving measurement accuracy. Accurate monitoring of vital parameters requires advanced image processing techniques combined with deep learning methods, which enable improved measurement precision, especially under challenging environmental conditions.

The traditional signal processing based methods used to estimate the pulse wave signal are fully explainable. However, usually the fixed masks for face, forehead or cheeks produce low quality signals. One of the key challenges in applying deep learning to medical contexts is its limited interpretability. Although these models can efficiently extract high-quality signals from input video, they often function as black boxes, raising concerns among medical professionals. In response to these issues, the field of Explainable Artificial Intelligence (XAI) is being developed, focusing on creating interpretable models while maintaining high predictive accuracy. One widely used solution is Grad-CAM [7], which allows for the visualization of image areas with the most significant impact on the model's decision. Grad-CAM has found applications in various aspects of facial analysis [8,9], including identity recognition, emotional expression analysis, attribute classification, and medical diagnostics [10,11].

This paper aims to combine both techniques by proposing a methodology that uses Grad-CAM to identify key image regions, enhancing pulse wave estimation accuracy through traditional signal analysis. The main assumption is that Grad-CAM-based masking enables the selection of the most relevant image areas, thereby increasing the SNR compared to analyzing entire images. To evaluate this approach, the Temporal Shift Convolutional Attention Network (TS-CAN) [17] was assessed using the rPPG-Toolbox [16] on the Pulse Rate Detection Dataset (PURE) [24]. Grad-CAM heatmaps were then generated to determine which facial regions had the most significant influence on the model's predictions, allowing for the identification of the most critical areas for rPPG signal extraction, and using these heatmaps, a binary mask was created. Additionally, a windowed approach to mask generation was introduced, incorporating varying numbers of neighboring frames in the computation of Grad-CAM heatmaps, and its impact on heart rate estimation accuracy was analyzed. The effectiveness of the proposed method was assessed using various metrics, including the SNR$_{raw}$, SNR$_{max}$, SNR$_{sum}$, and Hjorth descriptors. Moreover, statistical analyses were conducted to determine the significance of the results and to assess the influence of the applied methods on signal quality.

The rest of the paper is structured as follows. Section 2 provides a review of existing rPPG methodologies, with a particular focus on the impact of lighting conditions, ROI selection, and the application of deep neural networks in heart rate estimation. Section 3 describes the proposed methodology, including the implementation of Grad-CAM-based masking and the applied signal analysis techniques. Section 4 presents the experimental results and statistical analysis, followed by a discussion of the findings in Sect. 5. Finally, Sect. 6 concludes the paper by discussing potential improvements in rPPG signal acquisition.

2 Related Work

In recent years, rPPG has gained recognition as a progressive method for contactless monitoring of vital parameters. Research in this field focuses on identifying factors that affect measurement accuracy and improving signal processing methods. The study [12] analyzed the impact of various lighting conditions, frame rates, and video compression on the effectiveness of heart rate detection using videoplethysmography. It was shown that different ambient light sources significantly influence the quality of pulse estimation in rPPG, and optimal lighting conditions, along with proper camera settings, are crucial for achieving accurate measurements. The findings also emphasize the necessity of minimizing disturbances caused by video compression and further developing rPPG technology to address the challenges of real-world environmental conditions. The study [3] demonstrated the feasibility of measuring pulse rate using a webcam by analyzing color changes in specific facial regions, such as the forehead. The authors proposed focusing on selected areas of the face to enhance measurement accuracy and reduce the impact of disturbances. The study highlighted the effectiveness of the Principal Component Analysis (PCA) method in extracting the pulse signal while ensuring high computational efficiency, making it suitable for real-time applications. It was also noted that selecting smaller ROI helps minimize motion artifacts and provides reliable measurements under controlled lighting conditions. Further research [13] explored the use of normalized and tracked facial regions, such as the cheeks and nose, to improve signal stability and enhance the robustness of rPPG measurements. Recent advancements in rPPG research focus on leveraging deep neural networks to enhance the precision and reliability of pulse signal analysis. The study [14] demonstrated that the application of DNNs enables not only the classification of PPG signals based on their suitability for heart rate estimation but also the automatic identification of facial regions that provide the most stable measurements. Additionally, signal quality analysis, incorporating parameters such as SNR and Normalized Peak Energy (NPE), further improves classification accuracy and reduces the likelihood of erroneous readings. Unlike traditional methods such as ICA and PCA, which may produce false-positive results in the presence of noise, deep learning models demonstrate greater adaptability to varying lighting conditions while effectively minimizing motion artifacts. Furthermore, the use of DNN regression models allows for direct heart rate estimation without requiring complex filtering techniques, making this approach particularly promising for telemedicine applications and intelligent health monitoring systems.

One of the key aspects of modern rPPG methods is the utilization of attention mechanisms and deep learning architectures, as demonstrated in the study by [15]. The introduced MAR-rPPG method focuses on improving the precision of ROI localization and enhancing motion robustness. The implementation of masked attention enables increased semantic consistency in attention maps across consecutive video frames, while the masking technique prevents the model from overly relying on imprecisely determined ROIs. Additionally, the integration of the Enhanced rPPG Expert Aggregation (EREA) architecture allows for a dynamic analysis of various facial regions and selectively directs the model's attention to the most relevant features of the rPPG signal. The findings indicate that combining attention mechanisms with the elimination of incorrectly designated ROIs significantly enhances rPPG's resistance to motion artifacts, thereby improving the stability and accuracy of heart rate estimation. A comparable approach was adopted in [16], which introduced rPPG-Toolbox, a tool designed to assess and compare the effectiveness of different rPPG methods. The authors focused on deep learning models that employ convolutional neural networks and attention mechanisms for signal analysis.

Among the discussed architectures, TS-CAN was highlighted for its use of temporal shifts in signal analysis [17], PhysNet for its reliance on three-dimensional convolutional networks [18], DeepPhys for integrating attention mechanisms to select key features [19], and PhysFormer, which incorporates transformers for spatiotemporal analysis [20]. The application of these models enables more effective suppression of motion artifacts and improves heart rate estimation accuracy. Notably, the authors emphasize that the standardization of training processes and the analysis of signal quality impact on predictive performance contribute to the optimization of deep learning methods in rPPG. A significant feature of the rPPG-Toolbox is also its ability to integrate model interpretability mechanisms, such as Grad-CAM, which facilitates the identification of facial regions crucial for heart rate prediction [7].

An essential aspect of rPPG analysis is assessing signal quality, as it directly impacts the accuracy of heart rate estimation. The quality measure of rPPG signals is often represented by the value of the SNR. In many papers, the SNR quality measure for rPPG is defined as [21]:

$$SNR = 10log_{10}\frac{\sum_{f=30}^{240}(U_w(f)S_f)^2}{\sum_{f=30}^{240}((1-U_w(f))S_f)^2}, \tag{1}$$

where f denotes the frequency in bpm, $S(f)$ represents the spectrum of the extracted signal, $U_w(f)$ denotes the binary template window with the window size represented by w:

$$U_w(f) = \begin{cases} 1 & f_{PR} - \frac{w}{2} \leq f \leq f_{PR} + \frac{w}{2} \\ 1 & 2f_{PR} - \frac{w}{2} \leq f \leq 2f_{PR} + \frac{w}{2} \\ 0 & otherwise \end{cases} \tag{2}$$

and f_{PR} represents the true pulse rate value. The values f = 30 and f = 240 represent the lower and upper limits for possible pulse rates (0.5 Hz, 4 Hz). However, these frequencies are defined differently in some other papers (e.g., [22]). Using such a template window is a good proposition, assuming we know the reference value of the pulse rate. Some methods [23] propose extracting the pulse signal from image sequences using the SNR measure as an objective function during the training of the deep neural networks. This is a very interesting approach. However, there is uncertainty about whether false positive signals could be extracted to fulfill that criterion. The Hjorth descriptors were originally proposed for online analysis of EEG signals [25]. They are often used to represent the signal dynamics and purity in reference to the sinus function. These parameters will be used also in this study.

3 Method

In the following section, we provide details about the PURE dataset, including the experimental conditions, categorization of motion scenarios performed by participants, and the data selection process. We also describe the methodology for generating Grad-CAM masks and calculating signal quality metrics, such as SNR, Hjorth descriptors, which were used in the analysis.

3.1 Dataset

For the purpose of this study, data from the PURE dataset were utilized [24], including recordings of participants performing controlled head movements in front of a camera, while both facial images and reference heart rate measurements were captured. The participants were positioned approximately 1.1 m from the camera under natural daylight conditions. Each participant (01–10) engaged in six distinct scenarios, including: (1) remaining motionless, (2) speaking, horizontal head movements at different speeds – (3) slow and (4) fast translation, as well as looking at designated points around the camera, which induced (5) small and (6) medium head rotations ranging from 20° to 35°. A file naming convention $AA-BB$ was adopted, where AA represented the participant index and BB denoted the scenario number, e.g., $01-01$. During the recordings, participants' heart rates ranged from 42 BPM to 148 BPM, with all measurements obtained in a resting state. Due to issues with face detection during data preprocessing for TS-CAN inference and face mask generation, the sample for participant 05 and sample $06 - 02$ were removed from the analysis.

3.2 Explainability Analysis of the TS-CAN Model

The TS-CAN model was trained on the PURE dataset for contactless monitoring of physiological parameters, such as heart rate and respiratory rate, through video analysis. A fundamental component, the temporal shift module, efficiently models dynamic signal variations while maintaining low computational costs. This enables real-time operation on devices with limited processing power. Its architecture consists of two primary branches: the appearance branch, which detects subtle skin tone variations indicative of blood flow, and the motion branch, which captures micro-movements related to respiration and cardiac activity. Additionally, the attention module enhances signal to noise separation by focusing on the most relevant facial regions. The use of multi-task learning allows for the simultaneous estimation of heart rate and respiratory rate, optimizing computational efficiency and improving model accuracy.

The training process was conducted using rPPG-Toolbox, an open-source framework dedicated to rPPG signal analysis. The model was trained for 40 epochs with a batch size of 4 and a learning rate of $9 \cdot 10^{-3}$, achieving the best performance in epoch 34. Data from the PURE dataset underwent preprocessing, including DiffNormalization and Standardization. During initial processing, the facial region was detected in the first frame of each video, cropped, and expanded by 50% around the detected area. These regions were then resized to 72×72 px and used in both training and inference to generate Grad-CAM visualizations. In the conducted experiments, two masks were introduced: 1) mask m_1, which defines a rectangular facial region, and 2) mask m_2, derived from Grad-CAM heatmaps. First, using the MediaPipe detector and input frames in the RGB domain, employed for Grad-CAM computation, the facial region was identified, and a rectangular binary mask (m_1) was generated. Subsequently, Grad-CAM maps were computed for the Attention Mask 2 layer, normalized for each frame, and converted into binary masks using a 0.5 threshold. To enhance stability and facilitate comparative analysis, a windowed averaging approach was applied, where consecutive Grad-CAM maps were averaged to construct m_2. The window size was set to 1, 3, 5, 9, 13, 17, 21, 25, 29, and 33. For example, a mask size of 5 indicates that the Grad-CAM mask for a given frame is calculated using the two preceding, the target, and the two subsequent frames. When data from Attention Mask 2 were unavailable, the window size was progressively increased until the missing information was retrieved. The signal was computed based on the mean intensity of the green channel from input frames, multiplied by both masks. This resulted in two signal variants:

- s_1, where only the facial mask was applied,
- s_2, where both the Grad-CAM mask and the facial mask were applied simultaneously.

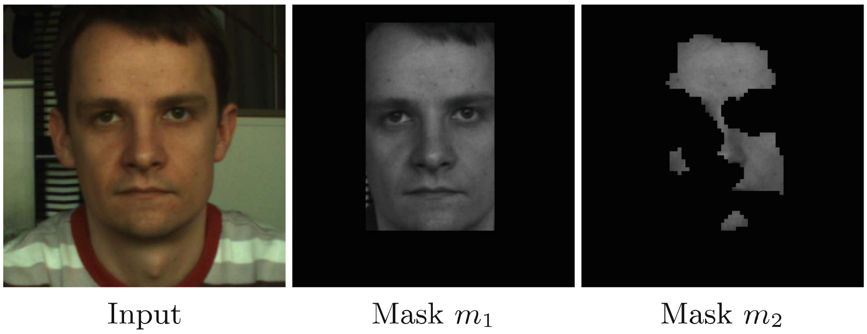

Fig. 1. Example of a single frame with different applied binary masks

Due to substantial variations in facial region size and the number of unmasked pixels across frames, masked pixels were excluded from signal computation. This prevented their influence on the computed averages, ensuring a more reliable signal comparison (Fig. 1).

3.3 Signal Quality Assessment

The next stage of the analysis involved computing SNR_{raw}, SNR_{sum} and SNR_{max} on the two extracted signals: s_1 and s_2. As part of the computation of the SNR_{raw} for the entire signal spectrum, a high-pass Butterworth filter was first applied with a cutoff frequency of 0.2 Hz, as lower frequencies were considered as baseline drift. Subsequently, a discrete Fourier transform (FFT) was performed on the filtered signal, resulting in the computed frequency spectrum $X(k)$, which was then analyzed in terms of its absolute value $|X(k)|$. The frequency range of [0.7,4.0] Hz was considered as the main component of the analyzed pulse signal, while frequencies outside this range were classified as a noise. The signal power, was determined by summing the squared values within the specified frequency range. Similarly, the noise power, was defined as the sum of squared values for frequencies outside the [0.7,4.0] Hz range. Finally, SNR_{raw} was defined using the (Eq. 1) formula, with the values $\mathrm{f}_{min} = 42$ and $\mathrm{f}_{max} = 240$ (beats per minute corresponding to 0.7–4 Hz) and $U_{wi}(f)$ (Eq. 3), where $U_{wi}(f)$ denotes the binary template window with the window size represented by w for the particular position of the template window i:

$$U_w(f) = \begin{cases} 1 & f - \frac{w}{2} \leq f \leq f + \frac{w}{2} \\ 0 & otherwise \end{cases} \tag{3}$$

SNR_{sum} (Eq. 4) and SNR_{max} (Eq. 5) measures refer to the maximum peak in the spectrum (describing the "energy" of the dominating frequency) without reference to the unknown, ground-truth pulse rate value. The values of these measures reflect how much some frequency (within the range of e.g., ±5 bpm, if w = 3) dominates in the spectrum, even if this fundamental frequency is not related to the pulse wave. This approach uses only information within measured signal in the potentially useful bandwidth, without unknown pulse rate (Eq. 1).

In our case, we assumed a frequency range of 0.7–4.0 Hz and a window size of 15 for the calculations.

$$SNR_{sum} = 10log_{10} \sum_{i=1}^{N-w} \frac{\sum_{f=42}^{240} (U_{wi}(f)S_f^2)^2}{\sum_{f=42}^{240} ((1 - U_{wi}(f))S_f^2)^2}, \tag{4}$$

$$SNR_{max} = 10log_{10}(max_{i=1}^{N-w} \frac{\sum_{f=42}^{240} (U_{wi}(f)S_f^2)^2}{\sum_{f=42}^{240} ((1 - U_{wi}(f))S_f^2)^2}) \tag{5}$$

Activity, mobility, and complexity Hjorth descriptors (Eq. 6) were adapted for the input signal y as:

$$\begin{gathered} activity = w_0 = \mathrm{var}(y), \quad mobility = \sqrt{\frac{w_2}{w_0}} = \sqrt{\frac{\mathrm{var}(\dot{y})}{\mathrm{var}(y)}}, \\ complexity = \sqrt{\frac{w_4}{w_0}} = \sqrt{\frac{\mathrm{var}(\ddot{y})}{\mathrm{var}(y)}}, \end{gathered} \tag{6}$$

where $\dot{y}$ and $\ddot{y}$ are the first and second derivative of the signal y. In the case of a discrete-time signal, activity is defined by the signal's variance, representing its total energy, while mobility describes the dominant frequency. The complexity (different than proposed by Hjorth) measures the high dynamics of the signal. For the analyzed domain of the signal these descriptors have lower values when variance of noise is low.

4 Results

The impact of different mask types and window sizes on signal quality was evaluated. The analysis was focused on computed SNR metrics and Hjorth descriptors, providing insights into performance differences between masks m_1 and m_2. The results were summarized in tables and figures, with Fig. 2 illustrating frames where the proposed Grad-CAM masks were applied.

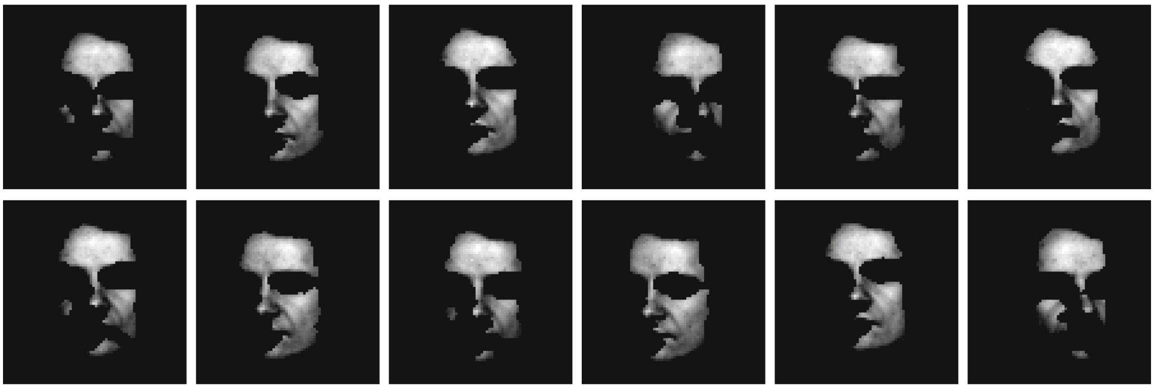

Fig. 2. Sequence of frames with extracted m_2 masks

4.1 Signal to Noise Ratio Analysis

The calculated SNR metrics for different mask types were presented in Table 1, which reported values for SNR$_{raw}$, SNR$_{max}$, and SNR$_{sum}$. The table included the differences between these metrics for m_1 and m_2.

Table 1. Comparison of SNR$_{raw}$, SNR$_{max}$, and SNR$_{sum}$ metrics

	m_1			m_2			m_2 - m_1	
	SNR$_{raw}$	SNR$_{max}$	SNR$_{sum}$	SNR$_{raw}$	SNR$_{max}$	SNR$_{sum}$	ΔSNR$_{raw}$	ΔSNR$_{max}$
1	0.1118	−0.9131	9.368	4.221	8.726	17.57	4.109	9.639
2	−0.6648	−8.272	4.549	−0.3398	−4.19	8.115	0.325	4.082
3	−2.592	−6.131	5.625	−0.1757	−2.208	8.505	2.416	3.923
4	−2.107	−5.653	5.975	0.1769	−6.374	5.749	2.284	−0.721
5	−5.082	−4.627	6.259	−1.886	1.646	11.91	3.196	6.273
6	−4.781	−5.7	5.963	−3.247	1.627	11.39	1.534	7.327

To further analyze the impact of mask types, Table 2 provided comparative results across all categories, incorporating Hjorth descriptors and their differences for both masks.

Table 2. Comparison of Hjorth descriptors: activity, mobility, and complexity

	m_1			m_2			m_2 - m_1		
	activity	mobility	complexity	activity	mobility	complexity	Δactivity	Δmobility	Δcomplexity
1	0.0001217	0.8115	1.334	4.277e-05	0.4664	0.6836	−7.4e-05	−0.3451	−0.6504
2	0.000247	0.612	0.9224	0.0001148	0.44	0.6078	−0.0001322	−0.172	−0.3146
3	0.001994	0.5858	0.9133	0.000993	0.5118	0.7366	−0.001001	−0.074	−0.1767
4	0.0004462	0.5629	0.8821	0.000166	0.5985	0.8774	−0.0002802	0.0356	−0.0047
5	0.0003806	0.4999	0.8056	8.876e-05	0.4134	0.6102	−0.0002918	−0.0865	−0.1954
6	0.0004796	0.4977	0.7921	0.0001968	0.3765	0.5532	−0.0002828	−0.1212	−0.2389

Table 3 summarized the statistical tests used to compare signals derived from m_1 and m_2, including the Mann-Whitney test (MW p-value and Rank Biserial Correlation), as well as the Shapiro-Wilk test results for groups $G1$ and $G2$. Additionally, the t-test, applicable under normal distribution assumptions, was included.

Table 3. Comparison of statistical metrics for SNR$_{raw}$ and SNR$_{max}$

	SNR$_{raw}$					SNR$_{max}$				
	MW p-val	MW RBC	SW G$_1$	SW G$_2$	TT p-val	MW p-val	MW RBC	SW G$_1$	SW G$_2$	TT p-val
1	0.0047	−0.8025	0.7928	0.9631	0.0012	0.0081	−0.7531	0.0293	0.2293	0.0044
2	0.8785	−0.0625	0.4058	0.2693	0.6107	0.2345	−0.3750	0.0453	0.3495	0.1353
3	0.0273	−0.6296	0.7643	0.2208	0.0274	0.0774	−0.5062	0.0282	0.5354	0.1322
4	0.0217	−0.6543	0.8759	0.7173	0.0164	0.8598	0.0617	0.3016	0.7395	0.7526
5	0.0423	−0.5802	0.2820	0.1148	0.0232	0.0171	−0.6790	0.1524	0.2598	0.0140
6	0.2893	−0.3086	0.2879	0.2378	0.2007	0.0020	−0.8765	0.1512	0.4680	0.0005

4.2 Window Size Comparison

The effect of window size on SNR$_{raw}$ and SNR$_{max}$ in the computation of the Grad-CAM binary mask was examined. Figures 3 and 4 displayed SNR$_{raw}$ for m_2 alongside baseline values for m_1. Figures 5 and 6 illustrated the differences in SNR$_{raw}$ and SNR$_{max}$, highlighting the impact of varying the window size on signal quality.

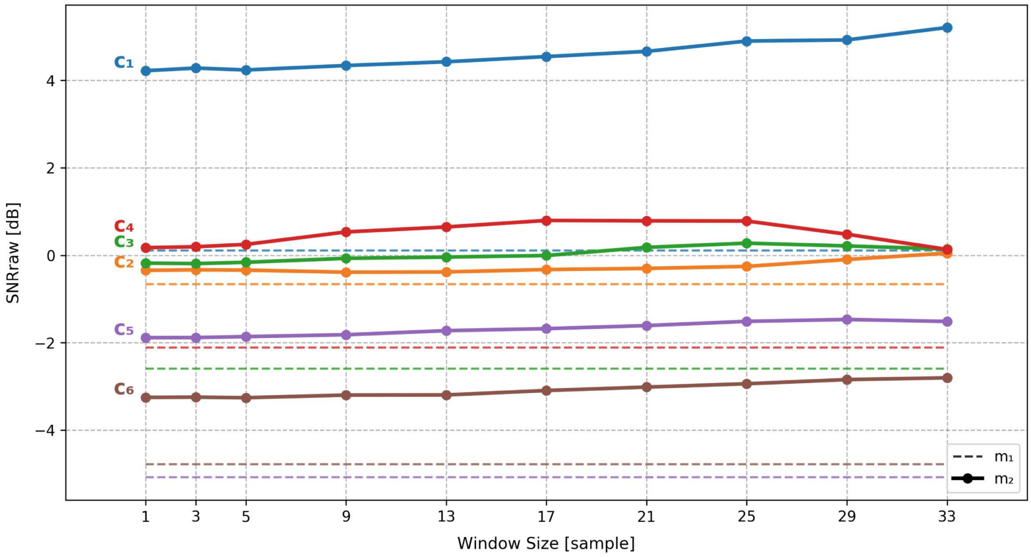

Fig. 3. Comparison of SNR$_{raw}$ for two masks across different window sizes

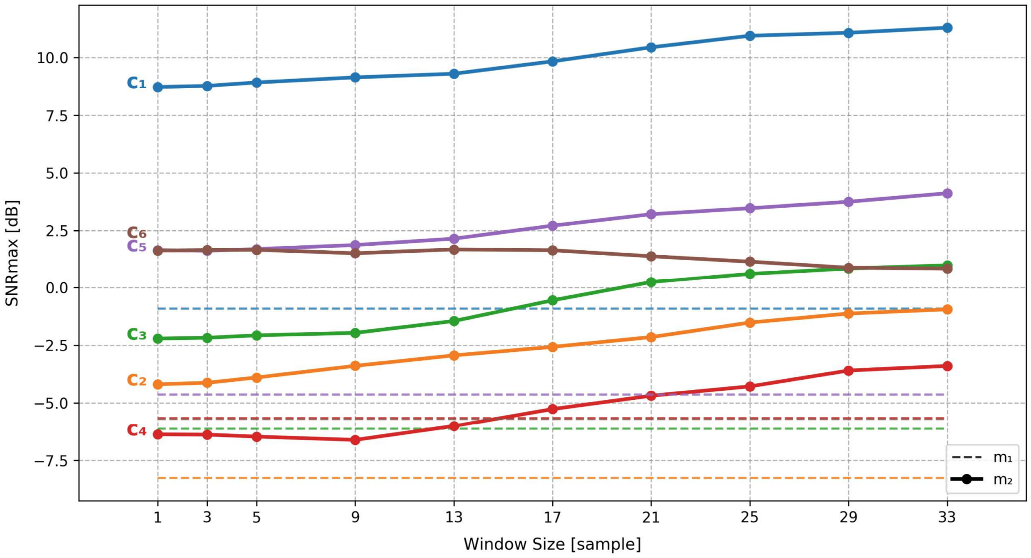

Fig. 4. Comparison of SNR$_{max}$ for two masks across different window sizes

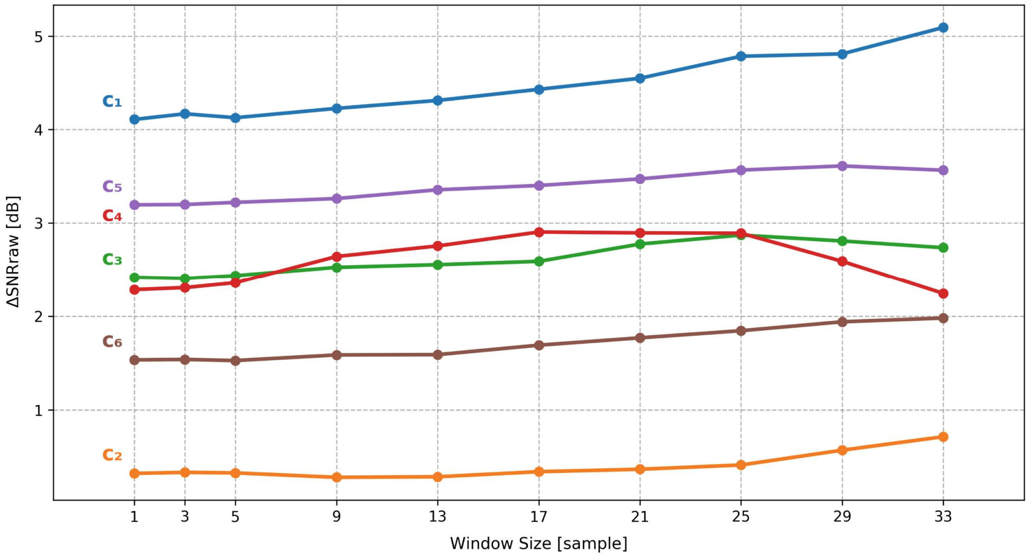

Fig. 5. Comparison of SNR$_{raw}(m_2)$ - SNR$_{raw}(m_1)$ across different window sizes

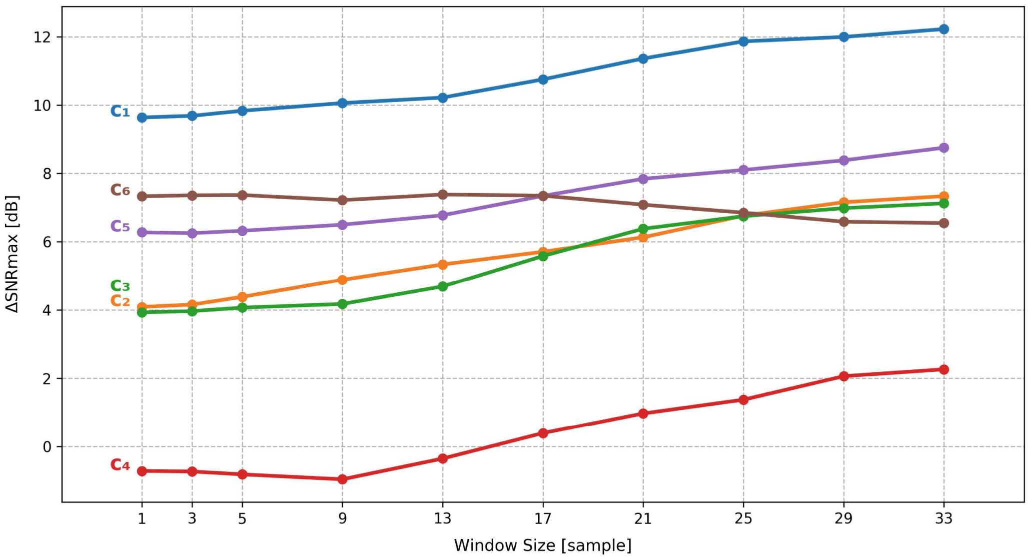

Fig. 6. Comparison of SNR$_{max}(m_2)$ - SNR$_{max}(m_1)$ across different window sizes

5 Discussion

The largest improvements in SNR$_{raw}$ with Grad-CAM-based masking were observed in category 1 (no movement), where the SNR increased by +4.109, and in category 5 (small head rotations), with an increase of +3.196. Notable increases occurred in categories 3 (slow head movement, +2.416), 4 (rapid head movement, +2.284), and 6 (large head rotations, +1.534). Category 2 (speaking activity) showed minimal improvement (+0.325), suggesting that speech artifacts persist despite masking. SNR$_{max}$ improved notably, with the largest increase in categories 1 (+9.639), 6 (+7.327) and 5 (+6.273). Moderate improvements were seen in categories 2 (+4.082) and 3 (+3.923), whereas category 4 showed a slight decline (−0.721), indicating potential distortions under fast motion using Grad-CAM masking. In our Hjorth descriptors implementation, all differences between masks m_2 and m_1 were negative, indicating reduced noise in the obtained signal, except for a slight mobility increase in one instance of category 4. Considering the temporal window size of the attention mask in relation to the increase in SNR$_{raw}$ and SNR$_{max}$, the findings indicate that SNR$_{raw}$ exhibits the greatest improvement in category 1 (+0.9840), while in categories 2, 3, and 5, the differences remain minimal, indicating that the expansion of the window led to only a slight improvement in signal enhancement. In contrast, SNR$_{max}$ shows a significantly greater improvement, particularly in categories 1, 2, 3, 4 and 5, where differences exceed +2, suggesting that larger window sizes substantially improved signal quality in the 0.7–4.0 Hz band.

Based on the Mann-Whitney and Shapiro-Wilk tests, significant differences between the analyzed groups were observed in certain categories, with both statistical significance (MW p-value) and effect size (MW RBC) playing a crucial role in result interpretation. For SNR$_{raw}$, significant differences were found in categories 1, 3, 4, and 5, as confirmed by the low MW p-values (MW p-val < 0.05) and large effect sizes ($|\text{MW RBC}| \geq 0.5$), confirming substantial differences between groups. Category 6, despite demonstrating a moderate effect size ($0.3 \leq |\text{MW RBC}| < 0.5$), did not reach statistical significance in the Mann-Whitney test, suggesting that the observed differences may be attributed to random data fluctuations rather than systematic variance. Meanwhile, category 2 exhibited neither statistical significance (MW p-val = 0.8785) nor a meaningful effect size ($|\text{MW RBC}| < 0.3$). For SNR_{max}, significant differences were observed in categories 1, 5, and 6, where (MW p-val < 0.05) and effect sizes were large ($|\text{MW RBC}| \geq 0.5$), indicating a strong distinction between the s_1 and s_2 signals. Category 2, despite reaching a moderate effect size ($0.3 \leq |\text{MW RBC}| < 0.5$), did not demonstrate statistical significance, while category 4 showed both a small effect size and no statistical significance. Category 3 exhibited a large effect size ($|\text{MW RBC}| = 0.5062$), but did not reach statistical significance.

This preliminary study has many limitations. First, it uses only one public dataset. Many other datasets can be analyzed in the future. Second, only one deep-learning model was used to extract Grad-CAM arrays and calculate masks used in this study. The TS-CAN model is very effective; however, more models have been proposed. Third, in our study, we considered only the signal from the green channel, whereas many traditional techniques, such as ICA and PCA, can significantly enhance signal quality by utilizing all three available channels. Nevertheless, this study presents a case showing how we can extract regions of a face that contribute most to the final, estimated BVP signal. It can be extended to propose a fully automated method to extract BVP-related face regions instead of fixed ROIs like the face area, forehead, cheeks, etc.

6 Conclusion

The performed analyses indicated that the application of Grad-CAM-based masking leads to a significant improvement in the quality of the rPPG signal. In addition to enhancing signal quality, the application of a binary mask derived from Grad-CAM, generated using the TS-CAN model for rPPG estimation, provides insight into the model's decision-making process, demonstrating that the identified regions contribute to the accurate extraction of blood volume pulse signals. In the majority of the analyzed scenarios, the SNR_{raw} and SNR_{max} values were significantly higher when using the Grad-CAM mask (m_2) compared to the reference face mask (m_1). Grad-CAM based masking provides the greatest benefit in stationary conditions and controlled movements. Moderate improvements occur for slow movements and speech, while fast head movements see no advantage or even slight degradation in SNR_{max}. Large head rotations benefit moderately but remain affected by motion artifacts. The analysis of the attention mask window size revealed that SNR_{raw} improves only marginally, while SNR_{max} increases by over 3.2 dB in categories 2, 3, and 4. The proposed approach indicates that integrating an attention module which leverages multiple temporal windows could further enhance the TS-CAN model.

References

1. Verkruysse W, Svaasand LO, Nelson JS (2008) Remote plethysmographic imaging using ambient light. Opt Exp 16(26):21434–45
2. Chen X, Cheng J, Song R, Liu Y, Ward RK, Wang Z (2019) Video-based heart rate measurement: recent advances and future prospects. IEEE Trans Instrum Meas 68:3600–3615
3. Lewandowska M, Ruminski J, Kocejko T, Nowak J (2011) Measuring pulse rate with a webcam-a non-contact method for evaluating cardiac activity. In: Federated conference on computer science and information systems (FedCSIS), Szczecin, Poland, pp 405–410. https://ieeexplore.ieee.org/document/6078233
4. Wang W, den Brinker AC, Stuijk S, de Haan G (2017) Algorithmic principles of remote PPG. IEEE Trans Biomed Eng 64:1479–1491
5. Ni A, Azarang A, Kehtarnavaz N (2021) A review of deep learning-based contactless heart rate measurement methods. Sensors 21(11):3719. https://doi.org/10.3390/s21113719
6. Cheng C-H, Wong K-L, Chin J-W, Chan T-T, So RHY (2021) Deep learning methods for remote heart rate measurement: a review and future research agenda. Sensors 21(18):6296. https://doi.org/10.3390/s21186296
7. Selvaraju RR, Cogswell M, Das A et al (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128:336–359. https://doi.org/10.1007/s11263-019-01228-7
8. Yang X, Wu B, Sato I, Igarashi T (2019). Directing DNNs Attention for Facial Attribution Classification using Gradient-weighted Class Activation Mapping. arXiv, abs/1905.00593
9. Park S, Wallraven C (2021) Comparing Facial Expression Recognition in Humans and Machines: Using CAM, GradCAM, and Extremal Perturbation. arXiv, abs/2110.04481
10. Xiao M, Zhang L, Shi W, Liu J, He W, Jiang Z (2021) A visualization method based on the Grad-CAM for medical image segmentation model. In: International conference on electronic information engineering and computer science (EIECS) 2021, pp 242–247
11. Chien J-C, Lee J-D, Hu C-S, Wu C-T (2022) The usefulness of gradient-weighted CAM in assisting medical diagnoses. Appl Sci 12(15):7748. https://doi.org/10.3390/app12157748
12. Przybyło J, Kańtoch E, Jabłoński M, Augustyniak P (2016) Distant measurements of plethysmographic signal in various lighting conditions using configurable frame-rate camera. Metrol Meas Syst 23:579–592
13. Tulyakov S, Alameda-Pineda X, Ricci E, Yin L, Cohn JF, Sebe N (2016) Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2396–2404
14. Ruminski J, Kwasniewska A, Szankin M, Kocejko T, Mazur-Milecka M (2019) Evaluation of facial pulse signals using deep neural net models. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), Berlin, Germany, pp 3399–3403. https://doi.org/10.1109/EMBC.2019.8857839
15. Zhao P et al (2024) Toward motion robustness: a masked attention regularization framework in remote photoplethysmography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops, pp 7829–7838

16. Liu X et al (2022) rPPG-Toolbox: Deep Remote PPG Toolbox. Neural Information Processing Systems
17. Liu X, Fromm J, Patel SN, McDuff DJ (2020) Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. arXiv, abs/2006.03790
18. Yu Z, Li X, Zhao G (2019) Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks. CoRR, abs/1905.02419. http://arxiv.org/abs/1905.02419
19. Chen W, McDuff D (2018) DeepPhys: video-based physiological measurement using convolutional attention networks. CoRR, abs/1805.07888. http://arxiv.org/abs/1805.07888
20. Yu Z, Shen Y, Shi J, Zhao H, Torr PHS, Zhao G (2021) PhysFormer: facial video-based physiological measurement with temporal difference transformer. CoRR, abs/2111.12082. https://arxiv.org/abs/2111.12082
21. de Haan G, Jeanne V (2013) Robust pulse rate from chrominance-based rPPG. IEEE Trans Biomed Eng 60(10):2878–2886. https://doi.org/10.1109/TBME.2013.2266196
22. Lin X (2014) Using blood volume pulse vector to extract rPPG signal in infrared spectrum. Master thesis, Eindhoven University of Technology. https://research.tue.nl/files/47000435/785066-1.pdf
23. Spetlik R, Franc V, Cech J, Matas J (2018) Visual heart rate estimation with convolutional neural network. In: Proceedings of british machine vision conference, pp 1–12
24. Stricker R, Müller S, Gross H-M (2014) Non-contact "video-based pulse rate measurement on a mobile service robot". In: Proceedings of 23st IEEE international symposium on robot and human interactive communication (Ro-Man 2014), Edinburgh, Scotland, UK. IEEE, pp 1056–1062
25. Sörnmo L, Laguna P (2005) EEG signal processing. In: Sörnmo L, Laguna P (eds) Bioelectrical signal processing in cardiac and neurological applications. Academic Press, pp 55–179. https://doi.org/10.1016/B978-012437552-9/50003-9