Camera-based Recovery of Cardiovascular Signals from Unconstrained Face Videos using an Attention Network

Yogesh Deshpande[1], Surendrabikram Thapa[2,3], Abhijit Sarkar[3], and A. Lynn Abbott[1]
[1]Bradley Department of Electrical and Computer Engineering, Virginia Tech, USA
[2]Department of Computer Science, Virginia Tech, USA
[3]Virginia Tech Transportation Institute, USA
{yogeshd, surendrabikram, asarkar1, abbott}@vt.edu

Abstract

This paper addresses the problem of recovering the shape morphology of blood volume pulse (BVP) information from a video of a person's face. Video-based remote plethysmography methods have shown promising results in estimating vital signs such as heart rate and breathing rate. However, recovering the instantaneous pulse rate signals is still a challenge for the community. This is due to the fact that most of the previous methods concentrate on capturing the temporal average of the cardiovascular signals. In contrast, we present an approach in which BVP signals are extracted with a focus on the recovery of the signal shape morphology as a generalized form for the computation of physiological metrics. We also place emphasis on allowing natural movements by the subject. Furthermore, our system is capable of extracting individual BVP instances with sufficient signal detail to facilitate candidate re-identification. These improvements have resulted in part from the incorporation of a robust skin-detection module into the overall imaging-based photoplethysmography (iPPG) framework. We present extensive experimental results using the challenging UBFC-Phys dataset and the well-known COHFACE dataset. The source code is available at https://github.com/yogeshd21/CVPM-2023-iPPG-Paper.

1. Introduction

Devices that perform convenient measurements of physiological signals have grown in popularity in recent years. For example, wearable devices by Fitbit [6], Apple [17], AliveCor [1], and others are capable of monitoring heart rate and other vital metrics. In addition to wearable devices, researchers have also considered the use of camera-based monitoring of physiological signals (e.g., [23, 31, 45, 46]). Circumstances such as the novel coronavirus pandemic have also increased awareness of benefits that can be obtained from convenient, noninvasive devices [24]. Unlike systems that require contact with the body, camera-based systems have the potential to be less intrusive in many situations like patient monitoring, and driver monitoring [21, 32, 38, 41]. Researchers have developed imaging-based systems that benefit from deep-learning techniques [4, 22]. However, more work is needed in sensing instantaneous (instance-level) physiological metrics.

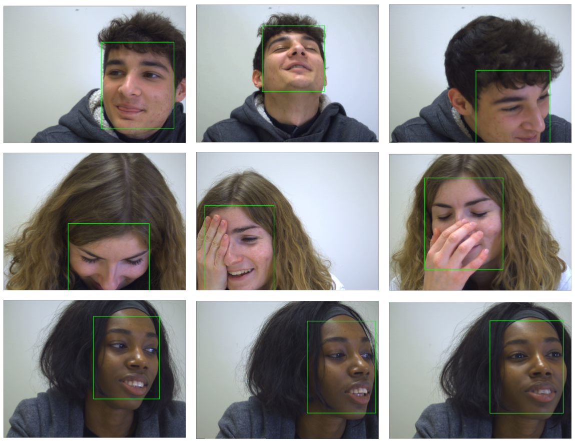

Figure 1. Examples of head movement and occlusion of faces in the UBFC-Phys dataset [30], highlighting our inclusion of natural movements. Green boxes indicate face regions detected by MTCNN [48].

This paper is concerned with monitoring the cardiovascular system through the analysis of image sequences from standard RGB video cameras. Sample frames are shown in Figure 1. The approach is based on the principle that each beat of the heart causes blood volume pulses (BVP) to travel through the body; these pulses cause slight changes in reflectance near the skin that are captured by the camera. The resulting intensity changes are very faint and are not noticeable with the unaided eye. The general technique is referred to in the literature as imaging-based photoplethysmography (iPPG) which makes use of light to perform remote measurements of volumetric changes.

While most previous approaches have focused on retrieving iPPG signals as the subject's head remains stationary during the measurement process, our work emphasizes the need to accommodate relatively large head movements. These movements pose significant challenges in detecting skin regions and measuring intensity changes accurately and reliably. Moreover, several confounding factors complicate the problem further, such as facial hair, eyeglasses, occlusion of the face, and variations in skin tone across subjects. Our work proposes a novel approach to address these challenges and improve the accuracy and reliability of camera-based recovery of cardiovascular signals in scenarios with significant head movements.

One of the key distinguishing features of our work is the focus on extracting individual BVP instances with good approximations of the underlying volumetric signal shape. Unlike previous systems that estimate average heart rate (HR), our approach has the potential to provide information related to inter-beat intervals and heart-rate variability (HRV). Another potential benefit of pulse-level signal information is the ability to distinguish one person from another. This paper also considers the problem of re-identification based on iPPG signals. We present a model that can make such re-identification possible. In summary, the primary contributions of this paper are as follows:

1) *New Architectural Pipeline:* The method relies on a deep network that incorporates a novel attention branch with a refined region of interest that emphasizes skin detection and also handles cases with facial hair and specular reflection, over a wide range of skin tones.

2) *Shape Morphology Recovery:* The new method emphasizes the recovery of shape morphology of physiological signals solely from RGB videos of the human face, with emphasis on handling large head movements and partial occlusion.

3) *Improved Standard Cardiovascular Metrics:* We present quantitative experimental results that demonstrate improved estimates of heart rate, as compared to previous state-of-the-art methods.

4) *Shape Morphology Metrics:* We present recovered time-variant physiological signal-based metrics and propose a standardized approach that could be followed by future researchers.

5) *Re-identification:* We introduce a candidate approach to subject re-identification, based on the recovered signals from the proposed model rather than averaged cardiovascular metrics. Our work therefore has the potential for use in biometric authentication tasks.

6) *Generalized ROI with Significant Variations:* Rather than handpicked regions of interest (ROI), our model targets generalization even in extreme cases including partial candidate visibility, occlusion cases, specular reflections from the skin as well as cases such as facial hair; all of them are addressed by our model.

2. Related Work

Approaches for contactless measurement of PPG signals and heart rate have been explored extensively over the past few years. Circumstances including the COVID-19 pandemic have led to significant changes in the healthcare sector [9, 18], including realization of the need for contactless techniques for the assessment of vital signs and physiological signals. An example is imaging-based measurement of PPG, which exploits signals from video of a subject to obtain BVP information. The subtle changes in skin pixel appearance due to blood flow is leveraged to estimate PPG signals and eventually vitals such as heart rate (HR) and heart rate variability (HRV). Previous research in the area can be broadly divided into three categories: signal processing-based methods, supervised methods, and unsupervised methods.

2.1. Signal Processing Based Methods

Using signal-processing techniques, the variation in average brightness of skin pixels is tracked over time [42]. This variation is too subtle to be noticed by human eyes without digital magnification [31]. Wu et al. [44] proposed a method to amplify such subtle changes. The method, commonly referred to in the literature as VidMag, takes a video sequence as input followed by temporal (band-pass) filtering of frames. The resulting signal after amplification was used to reveal hidden signals. Similarly, Garbey et al. [8] made use of sensitive thermal cameras to acquire the signals from major superficial vessels of face and neck regions. Fourier methods were used to measure cardiac pulse amplitudes. The main problem with these methods is the stability of the face in videos. The observed face is expected to remain stationary, and even small movements cause significant noise during PPG signal recovery. Later, researchers began utilizing a combination of signal-processing techniques and facial tissue trackers to tackle this problem [49].

2.2. Supervised Methods

With the increasing use of deep learning in the medical domain, supervised methods are now widely used for the retrieval of heart rate and physiological signals. Niu et al. [25] proposed an end-to-end HR estimation model based on spatial-temporal representations of multiple regions of interest (ROI). The use of multiple ROIs helped in the generalizability of the model in situations such as small movements, changing lighting conditions, etc. Similarly, DeepPhys [4] is a supervised model that makes use of a feed-forward CNN for the estimation of heart and breathing rates. DeepPhys incorporates an attention mechanism to assist the network in learning frame-to-frame differences. Earlier work relied primarily on high-resolution videos for the model implementation. Yu et al. [47] proposed a two-step mechanism for the estimation of heart rate from compressed videos. A video enhancement network is followed by an iPPG network to recover cardiovascular signals for estimating HR and HRV. The supervised model works well when there is a large number of training samples.

2.3. Unsupervised Methods

Because PPG signals are different for each individual, it can be difficult to learn general distributions. Thus, unsupervised methods have also been explored in recent years for the camera-based monitoring of PPG signals. For example, Lee et al. [14] proposed a meta-learning approach in a transductive setting for remote assessment of heart rate. Through this approach, substantial improvements were made in performances using the MAHNOB-HCI and UBFC-rPPG dataset [35]. Similarly, Wang et al. [40] proposed a self-supervised spatiotemporal learning framework for remote assessment of vitals such as HR using iPPG signals. A landmark-based spatial augmentation followed by sparsity-based temporal augmentation was used to cover diverse distributions in the approach. A constrained spatiotemporal loss was introduced to generate the pseudo-labels for augmented data. Unsupervised methods have also been used in sub-applications like skin tissue segmentation [2, 26].

2.4. Authentication

Every individual possesses a heart and associated vascular system that are inherently unique. When sensing physiological signals such as PPG and ECG, differences between individuals can lead to distinctive characteristics that can be leveraged for the purpose of biometric authentication [12, 33, 34]. Different research efforts have shown promise in the field of authentication using contact sensor-based PPG [15]. However, the development of biometric authentication systems using imaging PPG (iPPG) signals has been a challenging task due. One of the primary challenges is the presence of noise in iPPG signals, which can lead to inaccurate results. Additionally, the recovery of shape morphology from iPPG signals has been a difficult task for researchers [20, 27].

3. Architecture and Approach

A goal of this work is to accommodate relatively large movements of the head, and we wish to take advantage of Shafer's dichromatic reflection model [42]. We have modified the DeepPhys model [4] by incorporating a a skin segmentation model.

3.1. Extraction of Areas of Interest

The primary objective of this study is to incorporate natural variations in video data during the training process. Traditional face detection methods, such as center-crop, Haar cascades, and SeetaFace, are insufficient for locating faces with significant movement, partial occlusion, and wide ranges of skin tones. We decided to use the MTCNN face tracking and detection method [48], with results illustrated in Figure 1. This choice aided in addressing many problems related to face detection.

3.2. Skin Segmentation Model

Skin segmentation is a crucial step in training models for remote photoplethysmography (iPPG), as it directs learning toward the important areas of the face from where signals could be recovered. While previous research has emphasized the importance of skin pixels and proper skin detection, there has been less consideration given to the need for having a skin detection algorithm that would avoid facial hair, heavy skin illumination based reflection, glare, etc. Most existing ROI detectors do not have the capability to exclude these regions during learning. Additionally, when employing attention-based networks, it is often assumed that the regions of interest consist entirely of skin pixels.

To address these problems, we have used a skin segmentation model as shown in Figure 2. This is a fully convolutional network (FCN), with an encoder-decoder architecture. The model was trained using the benchmark ECU dataset [28] using binary cross entropy loss function and SGD optimizer. where the loss function could be represented as follows,

$$L_{Skin} = -1/N \sum_{i=1}^{N} y_i \log(f(y_i) + (1 - y_i) \log(1 - f(y_i)) \quad (1)$$

where N represents the number of classes (here $N = 2$ for skin pixels and non-skin pixels), y_i represents labels, and $f(y_i)$ represents predicted probability.

This trained model gives a skin probability mask as the output which is used in our main architecture to gain the output skin frame. To retrieve the skin frame, a thresholding operation is performed on the mask generated from the FCN model such that 0 represented non-skin pixels and 1 represented skin pixels. This computed mask is then multiplied with the RGB channels of the original face frame, thus providing the required skin ROI from the respective input video frame. Standard data augmentation techniques were used during training, along with image color variations in order to accommodate different skin tones.

The primary objective of this implementation is to focus on skin regions and eliminate areas that include signif-

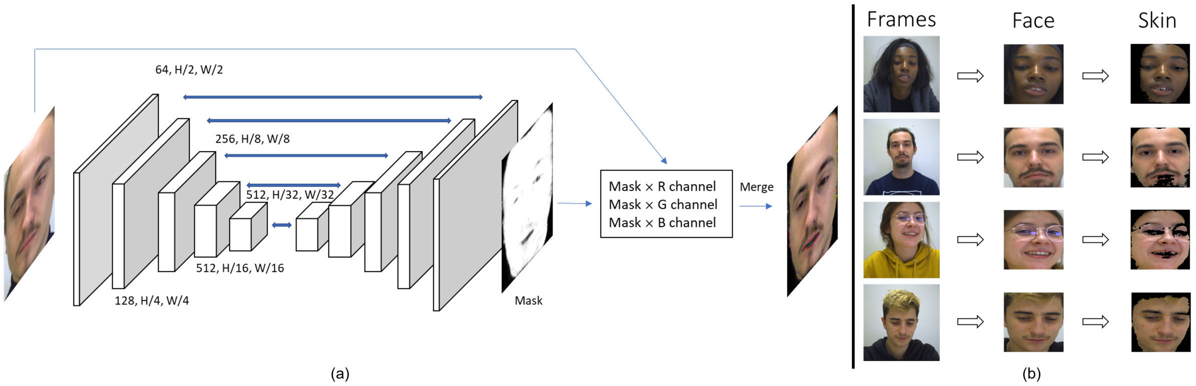

Figure 2. (a) Our skin segmentation model is a fully convolutional network (FCN) based encoder-decoder network. This subsystem generates a face mask in which skin pixels are detected. (b) Example outputs of some of the important cases using our ROI pipeline. The system extracts the face region and detects skin while excluding major facial hair and divergent factors such as reflections from eyeglasses.

icant facial hair, overly bright intensities, and specular reflections. The architecture of the model is presented in Figure 2(a). Figure 2(b) illustrates some critical test cases from the UBFC-Phys dataset along with their corresponding skin segmentation results, providing an intuition of how disregarding such scenarios could lead to a lack of generality during model training.

3.3. Proposed Model

Our proposed architecture is based on Shafer's dichromatic reflection model (DRM) [42] and the mathematical analogy introduced by DeepPhys [4]. We can represent the time-varying function of the RGB values of the k^{th} skin pixel in an image sequence as follows,

$$C_k(t) = I(t) \cdot (v_s(t) + v_d(t)) + v_n(t) \tag{2}$$

where $C_k(t)$ represents a vector of the RGB values, $I(t)$ represents luminance intensity level, $v_s(t)$ represents specular reflection, $v_d(t)$ represents diffuse reflection and $v_n(t)$ represents camera quantization noise. Here to reduce the effects due to camera quantization error, every frame is down-sampled to a size preferred by the model (72 × 72 in our case). Then bilinear interpolation for downsampling is used. This is in contrast to more conventional bicubic interpolation, as the former helps in avoiding excessive smoothing effects and influences better learning from the face features for the BVP signal retrieval. A revised representation is

$$C_l(t) = I(t) \cdot (v_s(t) + v_d(t)) \tag{3}$$

where $C_l(t)$ represents a vector of the RGB values for the l^{th} skin pixel from the resized frames. In previous modeling approaches, the l^{th} pixel in an image sequence is naively assumed to be a skin pixel. Instead, we have added an additional attention branch in the training pipeline (see Figure 3) that can improve the automatic selection of regions of interest to skin areas. As a result, we can update (3) as

$$C_l(t) = \begin{cases} 0, & \text{if } Skin(l) < \delta, \\ I(t) \cdot (v_s(t) + v_d(t)), & \text{if } Skin(l) \geq \delta \end{cases} \tag{4}$$

where $Skin(l)$ represents the outcome of our skin detection model for the l^{th} pixel and δ represents the threshold for the binary cross entropy probabilities (0.5 in our case).

We use MTCNN-based face detection and tracking [48] to extract target face regions from video frames. This choice helps our system accommodate spatial motion. We use a normalized face frame difference of two consecutive face frames as the input to the main branch of our convolution attention network (CAN), with maxima that are clipped to the third standard deviation above the mean. The input normalized face frame difference can be represented as

$$D_{ip}(t) = min(D(t), D_{ipmax}) \tag{5}$$

and

$$D(t) = \frac{C_l(t+1) - C_l(t)}{C_l(t+1) + C_l(t)} \tag{6}$$

$$D_{ipmax}(t) = \mu(D(t)) + (3 \times \sigma(D(t))) \tag{7}$$

where $C_l(t)$ is the vector of the RGB values for the resized face frame, $D_{ip}(t)$ is the input (clipped normalized face frame difference), $D(t)$ is the normalized face frame difference without clipping, $D_{ipmax}(t)$ is the maxima of the threshold for clipping, μ represents mean and σ represents standard deviation.

The attention branch of our model (which includes the skin segmentation model) helps retrieve the skin regions

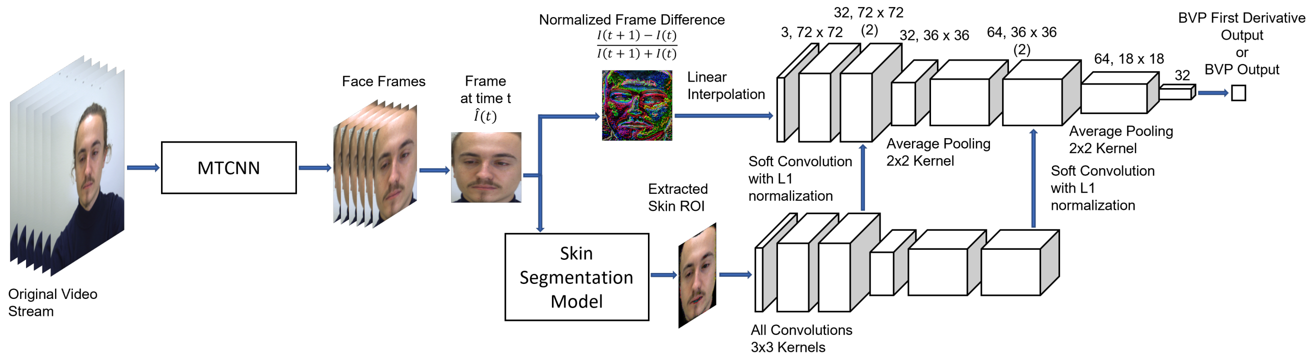

Figure 3. The complete proposed architecture. Face extraction is performed by the MTCNN module, and skin detection is performed on these areas of interest to recover the required ROI's. From a single frame difference, the system generates a single signal value either for a BVP signal or for a temporal derivative of the BVP signal.

from the face which are then batch standardized and passed on to the network. As depicted in Figure 3 the architecture after the skin segmentation model, in the attention branch, is the same as the one we have in the main branch of the CAN network. We use dropout layers, with dropout rates of 0.5, before every average pooling layer and also before the last fully connected layer which is followed by a tanh activation function. The mask generated from our attention branch uses sigmoid activation over the respective branch outcome which is multiplied by the height and width of the respective layer prior to pooling layers and then this outcome is divided by twice the L_1 normalization on the output of the sigmoid activation. Finally, for our feature extracting dense layers we consider 32 feature parameters in the output layer, which has a tanh activation function to keep the outcomes bounded. We trained our attention model using the Stochastic Gradient Descent with Momentum (SGDM) optimizer, a momentum of 0.9, a batch size of 128, and a learning rate of 10^{-4}.

3.4. BVP Annotation and Loss Function

During training, the ground truth blood volume pulse (BVP) values are first resampled to match the sampling rate of the video frames. The first derivative of these BVP signals is computed and batch standardized to be used as the ground truth in one training pipeline. We also use the original batch-standardized BVP signal as the ground truth, thus computing outcomes on both the first derivative as well as the original BVP signals in two different training pipelines. To compute the loss during training, we utilize the mean square error between the model outcome and the standardized ground truth. Hence, in the case of the first derivative BVP signal retrieval, it could be represented as

$$b_{der}(t) = b(t+1) - b(t) \tag{8}$$

$$b_{gt}(t) = \frac{b_{der}(t) - \mu(b_{der}(t))}{\sigma(b_{der}(t))} \tag{9}$$

$$Loss_{CAN} = \frac{1}{N}\sum_{i=1}^{N}(b_{gt}(t) - b_{pred}(t))^2 \tag{10}$$

where $b(t)$ is the BVP value collected from the sensor at time t, $b_{der}(t)$ is the first derivative signal, $b_{gt}(t)$ is the standardized first derivative BVP signal that we use as the ground truth and $b_{pred}(t)$ is the predicted model outcome.

3.5. Signal Morphology

The field of remote photoplethysmography (iPPG) has mainly focused on extracting average cardiac pulse-based metrics. However, as physical sensor-based technology advances, the potential for generating instantaneous physiological data also increases, highlighting the need for more research in this area [19]. This work is one of the first to tackle the challenges of detailed shape (morphological) features. In this section, we present a set of metrics that can be used to study conformity of the recovered signals.

For morphology-based metrics, we compute the mean of the normalized cross-correlation between the model output signals and the ground truth BVP signals for every candidate in the dataset. These metrics are computed for the respective signals in the time domain, frequency domain as well as power domain, which are reported further giving us a complete idea of how well the model could retrieve correct signal shape morphology.

The normalized cross-correlation (ncr) is computed as:

$$ncr(x_{gt}(n), x_{op}(n)) = \frac{\sum_{i=1}^{N} x_{gt}(n_i)x_{op}(n_i)}{\sqrt{\sum_{i=1}^{N} x_{gt}(n_i)^2}\sqrt{\sum_{i=1}^{N} x_{op}(n_i)^2}} \tag{11}$$

where $x_{gt}(n)$ is the ground truth signal, $x_{op}(n)$ is the model output signal and N is the number of signal samples.

The signal in the time domain is represented as $x_{gt}(t)$ and $x_{op}(t)$ where $x_{gt}(t)$ is ground truth signal in time domain and $x_{op}(t)$ is model output signal in time domain.

Thus, the same signals in the frequency domain could be represented as follows,

$$x_{gt}(f) = FFT_{mag}(x_{gt}(t)) \tag{12}$$

$$x_{op}(f) = FFT_{mag}(x_{op}(t)) \tag{13}$$

where FFT_{mag} is the magnitude of the Fast Fourier Transform for a signal in the time domain.

Similarly, the signals in the power domain will be as follows,

$$psd(x(n), f_s) = \lim_{x \to \infty} \frac{1}{T} \left| \sum_{n=1}^{N} x_n e^{-i2\pi f n} \right|^2 \tag{14}$$

$$x_{gt}(p) = psd(x_{gt}(t), f_s) \tag{15}$$

$$x_{op}(p) = psd(x_{op}(t), f_s) \tag{16}$$

where psd is power spectral density, f_s is sampling frequency and N is the number of signal samples.

Thus, based on this developed baseline we further compute our shape morphology metrics in the time, frequency, and power domain denoted as smm_t, smm_f and smm_p respectively, which could be given as follows:

$$smm_t = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(t)_{gt}, x_i(t)_{op}) \tag{17}$$

$$smm_f = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(f)_{gt}, x_i(f)_{op}) \tag{18}$$

$$smm_p = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(p)_{gt}, x_i(p)_{op}) \tag{19}$$

4. Experiments and Results

Since our implementation addresses the process of BVP signal recovery from face videos as well as re-identification based on those recovered signals, we divided our experiments into three parts, starting with the signal recovery forefront, followed by computation of standard human understandable metrics and then computing re-identification procedure. For better evaluation, we have two different pipelines covering both the recovery from ground truth BVP signals as well as first derivative BVP signals.

4.1. Dataset

UBFC-Phys [30]: This dataset includes 56 candidates with a distribution of 10 males and 46 females. The video frame rate is 35 FPS and has a resolution of 1024×1024, BVP signals are also provided in this dataset which are collected at a rate of 64 Hz using the E4 wristband. Each candidate is subjected to 3 tasks, which are rest, speech, and arithmetic tasks, respectively, and a 3-minute RGB video is collected for each task. All the videos were captured in a lab environment, with natural movements incorporated into all tasks. Additionally, the dataset features natural translational and rotational movements, along with various attributes such as facial hair, glasses, skin color, and occlusion. The annotated labels in the UBFC-Phys dataset for tasks two and three are not suitable for training, and hence we have only considered the data from the first task, which still encompasses all the necessary variations in the video data.

COHFACE [11]: This dataset includes 40 candidates with a distribution of 28 males and 12 females. The videos were captured at a frame rate of 20 FPS and have a resolution of 640×480. The dataset also includes corresponding BVP signals collected at a rate of 256 Hz. Each candidate in the dataset was recorded for a duration of one minute, under two different illumination scenarios, i.e., good lighting and low light conditions. All the videos were collected in a lab environment. Though there are no intentional movements, the challenge in this dataset is brought by the low illumination samples. The data also include a considerable variation of skin tones which makes it even more useful.

4.2. Signal Morphology Recovery

This section reports results on the recovery of signal shape morphology. Here, we demonstrate our model's performance on the UBFC-Phys dataset quantatively using (17)-(19) as well as using a visual depiction as shown in Figure 4. Table 1 presents the shape morphology metric outcomes for our models with ground truth as the original BVP signal and first derivative BVP signal, along with recovered signals from our implementation of the DeepPhys model all without any form of the post-processing involved. Similarly, in Table 2, we present the shape morphology metrics on the integrated signals after post-processing for the first derivative ground truth BVP signal-based models.

4.3. Standard Cardiac Pulse Metrics

The computation of the heart rate in beats per minute, is often performed by post-processing the model's output BVP signals using a bandpass filter. We have used a cutoff frequency of 0.75 Hz and 0.25 Hz (since the expected range for heart rate is 45 beats/min to 150 beats/min). We next compute the power spectral representation of the bandpassed signal where the highest peak is considered as the

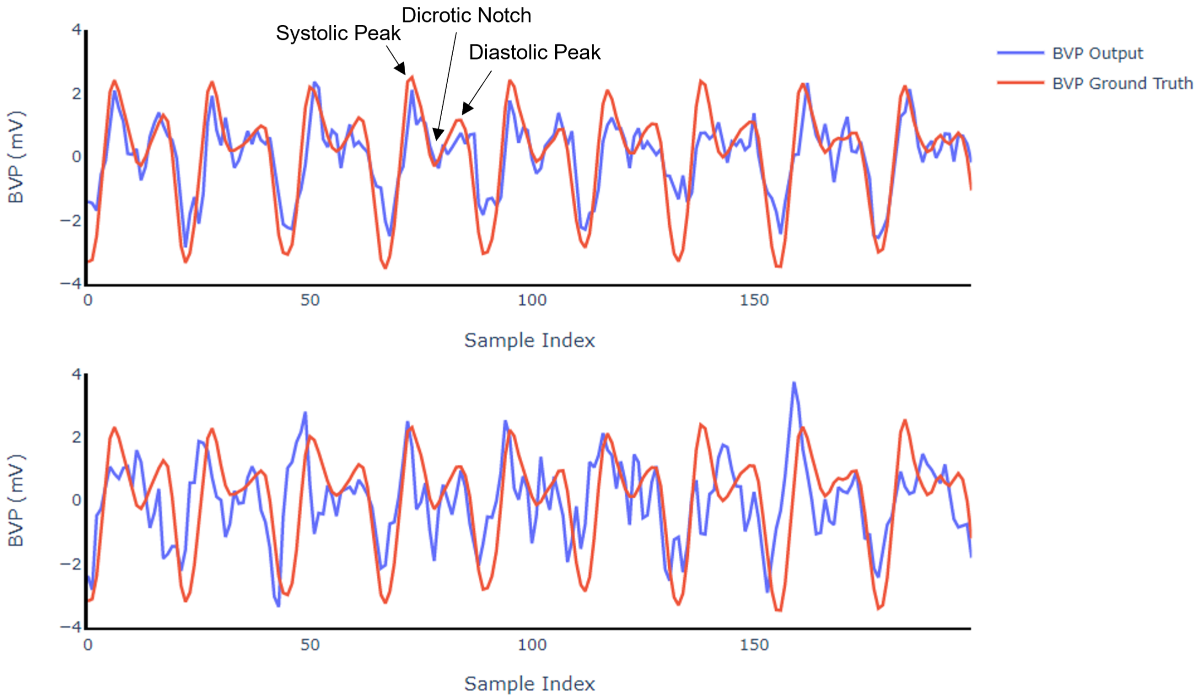

Figure 4. The two plots are for the same individual from the UBFC-Phys dataset, representing the shape morphology recovery from our model (top) and from DeepPhys [4] (bottom). This is a qualitative representation of how well our model retrieves signal shape morphology and its comparison with signal recovery from a state-of-the-art model that focuses on averaged pulse values.

Table 1. Domain-wise shape morphology metrics outcomes for our model pipelines and for DeepPhys without any post-processing of the output signals.

Metrics	Our Model (BVP)	Our Model (First Der. BVP)	DeepPhys
Time↑	**0.088**	0.077	0.06
Frequency↑	0.443	**0.511**	0.359
Power↑	0.45	**0.47**	0.339

Table 2. Domain-wise shape morphology metrics outcomes for our BVP first derivative-based model pipeline and DeepPhys after integrating output signals to get them in original BVP signal format.

Metrics	Our Model (First Der. BVP)	DeepPhys
Time↑	**0.120**	0.118
Frequency↑	**0.670**	0.645
Power↑	**0.594**	0.472

estimated HR. We report root mean square error (RMSE), mean absolute error (MSE), and Pearson correlation coefficient between the heart rate for the ground truth BVP signal and the estimated BVP signal. The results are shown in Table 3 for both UBFC-Phys and COHFACE.

Further, in Table 4 we present a comparison of the MAE-based outcomes for the average cardiac pulse-based measurements from our models with respect to the state-of-the-art models previously published.

4.4. Re-identification

As a part of our experiments, our aim was to evaluate the possible scope and extent of re-identification using our devised architecture. Since we were focusing primarily on good BVP signal shape retrieval and thereby performing computations based on the retrieved BVP signals, hence instead of evaluating authentication based on an Inter Beat Interval (IBI) or similar averaged cardiac pulse-based metrics, we computed the Pearson correlation coefficient between the ground truth and the output BVP signals for each candidate. So, the outcome for every candidate was compared with the annotated BVP signals for every other candidate in the test set and the one with the maximum correlation was acknowledged as the identified candidate for the respective output signal. Considering rank 5, we could re-identify 14 candidates from a pool of 20 candidates from a diverse dataset such as UBFC-Phys. The rank-5 accuracy was therefore 70%, which demonstrates the poten-

Table 3. Performance of our architecture pipelines on UBFC-Phys and COHFACE dataset, in terms of heart rate measurements in beats per minute (HR bpm) [7, 10, 13, 36]. Comparisons have been made with literature using available metrics that are used in our study.

Methods	UBFC-Phys				COHFACE			
	MAE↓	RMSE↓	r↑	$\overline{SNR}$(dB)↑	MAE↓	RMSE↓	r ↑	$\overline{SNR}$(dB)↑
ICA [29]	6.71	-	-	-	12.24	15.67	0.24	-4.43
CHROM [5]	4.39	-	-	-	7.80	12.45	0.26	-
POS [42]	5.98	-	-	-	13.43	17.05	0.24	-4.43
HR-CNN [36]	-	-	-	-	8.10	10.78	0.29	-
DeepPhys [4]	11.78	17.848	0.174	-7.676	6.60	10.788	0.524	-6.425
Our Model (BVP)	5.02	10.673	0.701	-1.792	4.02	6.799	0.80	-6.317
Our Model (1^{st} Der. BVP)	**4.05**	**8.438**	**0.828**	**-0.78**	**2.92**	**6.128**	**0.86**	**-2.685**

Table 4. Performance (HR BPM-MAE) of our technique in comparison with previously published state-of-the-art models on the UBFC-Phys and the COHFACE dataset [7, 10, 13, 36].

Methods	**UBFC Phys**	**COHFACE**
GREEN [39]	14.17	-
ICA [29]	6.71	12.24
CHROM [5]	4.39	7.8
POS [42]	5.98	13.43
1D-CNN [36]	5.41	-
LSTM-rPPG [3]	6.48	-
SQA-rPPG [7]	6.01	-
2SR [43]	-	20.98
LiCVPR [16]	-	19.98
HR-CNN [36]	-	8.10
SAMC [37]	-	6.23
DeepPhys [4]	11.78	6.6
Our BVP	5.02	4.02
Our BVP Der.	**4.05**	**2.92**

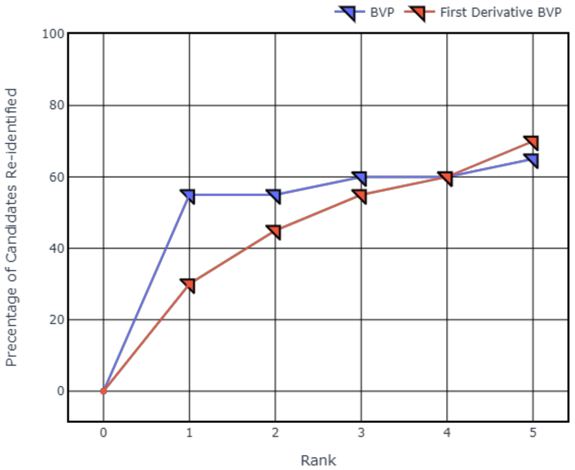

Figure 5. The rank-wise distribution for re-identification is presented here, where the graph in blue represents the re-identification results for the model trained using BVP signals, and the graph in red represents the re-identification results for the model trained using first derivative BVP signals.

tial of this approacy. The rank-wise distribution is presented in Figure 5 where we show the rank-wise re-identification for both of our models including first derivative BVP signal outcomes as well as integrated BVP signal outcomes.

5. Conclusion

This paper has presented an iPPG method that can extract BVP signals from standard RGB video of a person's face. The primary emphasis has been to recover the shape (morphology) of the BVP signal. We have shown that recovery of systolic and diastolic peaks is possible through camera-based iPPG. Using large-scale benchmark datasets and a series of metrics, we have demonstrated that our method performs better than previous state of the art methods to extract the BVP signal.

A better understanding of BVP will help iPPG research in many ways. First, there is no longer a need to place so much emphasis on recovering average heart rate only. Direct BVP signal recovery will help in studying inter-beat intervals with greater accuracy than is now possible. In turn, this work opens up the potential of iPPG in performing measurements related to heart rate variability. We have shown that better recovery of BVP signals significantly reduces error associated with other HR metrics. Finally, we also demonstrated that extracted BVP signals can be used for person reidentification.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 2136915.

Our project page can be visited to stay up-to-date on any developments regarding this work. A novel dataset collected with two NIR and one RGB camera will also be made available on our project page: https://sites.google.com/view/vt-tricam-ppg

References

[1] AliveCor, Inc. *AliveCor for Physiological Measurements*. Web page: https://www.alivecor.com. Accessed January 2023. 1

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 3

[3] Deivid Botina-Monsalve, Yannick Benezeth, Richard Macwan, Paul Pierrart, Federico Parra, Keisuke Nakamura, Randy Gomez, and Johel Miteran. Long short-term memory deep-filter in remote photoplethysmography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 8

[4] Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 1, 2, 3, 4, 7, 8

[5] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 8

[6] Fitbit, Inc. *Fitbit Watch for Physiological Measurements*. Web page: https://www.fitbit.com. Accessed January 2023. 1

[7] Haoyuan Gao, Xiaopei Wu, Jidong Geng, and Yang Lv. Remote heart rate estimation by signal quality attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2122–2129, 2022. 8

[8] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54(8):1418–1426, 2007. 2

[9] Ellen Gorman, Bronwen Connolly, Keith Couper, Gavin D. Perkins, and Daniel F. McAuley. Non-invasive respiratory support strategies in COVID-19. *The Lancet Respiratory Medicine*, 9(6):553–556, 2021. 2

[10] Amogh Gudi, Marian Bittner, and Jan Van Gemert. Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation. *Applied Sciences*, 10(23):8630, 2020. 8

[11] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017. 6

[12] Rudi Hoekema, Gérard J.H. Uijen, and Adriaan Van Oosterom. Geometrical aspects of the interindividual variability of multilead ECG recordings. *IEEE Transactions on Biomedical Engineering*, 48(5):551–559, 2001. 3

[13] Min Hu, Dong Guo, Xiaohua Wang, Peng Ge, and Qian Chu. A novel spatial-temporal convolutional neural network for remote photoplethysmography. In *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2019. 8

[14] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rPPG: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, pages 392–409. Springer, 2020. 3

[15] Lin Li, Chao Chen, Lei Pan, Jun Zhang, and Yang Xiang. SoK: an overview of PPG's application in authentication. *arXiv preprint arXiv:2201.11291*, 2022. 3

[16] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4264–4271, 2014. 8

[17] G. Lin, T. Nakajima, P. Rahul, and A. Hodge. *Seamlessly embedded heart rate monitor*. U.S. Patent 8,615,290. 1

[18] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2

[19] Xin Liu, Shwetak Patel, and Daniel McDuff. Camera-based physiological sensing: Challenges and future directions. *arXiv preprint arXiv:2110.13362*, 2021. 5

[20] Giulio Lovisotto, Henry Turner, Simon Eberz, and Ivan Martinovic. Seeing red: PPG biometrics using smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3

[21] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. SparsePPG: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1272–1281, 2018. 1

[22] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):1–40, 2023. 1

[23] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960, 2014. 1

[24] Seyedfakhreddin Nabavi and Sharmistha Bhadra. Design and development of a wristband for continuous vital signs monitoring of COVID-19 patients. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6845–6850. IEEE, 2021. 1

[25] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 2

[26] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3

[27] Omkar R. Patil, Wei Wang, Yang Gao, Wenyao Xu, and Zhanpeng Jin. A non-contact PPG biometric system based on deep neural network. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2018. 3

[28] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification:

analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005. 3

[29] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010. 8

[30] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. UBFC-Phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021. 1, 6

[31] Abhijit Sarkar. *Cardiac signals: remote measurement and applications*. PhD thesis, Virginia Tech, 2017. 1, 2

[32] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Assessment of psychophysiological characteristics using heart rate from naturalistic face video data. In *IEEE International Joint Conference on Biometrics*, pages 1–6, 2014. 1

[33] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. ECG biometric authentication using a dynamical model. In *IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015. 3

[34] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Biometric authentication using photoplethysmography signals. In *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016. 3

[35] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011. 3

[36] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, U.K.*, pages 3–6, 2018. 8

[37] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe. 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*, pages 2396–2404, 2016. 8

[38] Rik van Esch, Kambez Ebrahimkheil, Iris Cramer, Wenjin Wang, Tomas Kaandorp, Federica Sammali, A. T. M. Dierick-van Daele, Carla Kloeze, Cindy Verstappen, Marcel van't Veer, et al. Remote PPG for heart rate monitoring: lighting conditions and camera shutter time. In *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021. 1

[39] Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008. 8

[40] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2431–2439, 2022. 3

[41] Wenjin Wang and Albertus C. den Brinker. Camera-based respiration monitoring: Motion and PPG-based measurement. In *Contactless Vital Signs Monitoring*, pages 79–97. Elsevier, 2022. 1

[42] Wenjin Wang, Albertus C. Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 2, 3, 4, 8

[43] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974–1984, 2015. 8

[44] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):1–8, 2012. 2

[45] Umang Yadav, Sherif N. Abbas, and Dimitrios Hatzinakos. Evaluation of PPG biometrics for authentication in different states. In *2018 International Conference on Biometrics (ICB)*, pages 277–282, 2018. 1

[46] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021. 1

[47] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019. 3

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1, 3, 4

[49] Yan Zhou, Panagiotis Tsiamyrtzis, Peggy Lindner, Ilya Timofeyev, and Ioannis Pavlidis. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5):1280–1289, 2012. 2