A Temporal Encoder-Decoder Approach to Extracting Blood Volume Pulse Signal Morphology from Face Videos

Fulan Li[1], Surendrabikram Thapa[2,3], Shreyas Bhat[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
{lfulan, surendrabikram, bshreyas, asarkar1, abbott}@vt.edu

Abstract

This paper considers methods for extracting blood volume pulse (BVP) representations from video of the human face. Whereas most previous systems have been concerned with estimating vital signs such as average heart rate, this paper addresses the more difficult problem of recovering BVP signal morphology. We present a new approach that is inspired by temporal encoder-decoder architectures that have been used for audio signal separation. As input, this system accepts a temporal sequence of RGB (red, green, blue) values that have been spatially averaged over a small portion of the face. The output of the system is a temporal sequence that approximates a BVP signal. In order to reduce noise in the recovered signal, a separate processing step extracts individual pulses and performs normalization and outlier removal. After these steps, individual pulse shapes have been extracted that are sufficiently distinct to support biometric authentication. Our findings demonstrate the effectiveness of our approach in extracting BVP signal morphology from facial videos, which presents exciting opportunities for further research in this area. The source code is available at https://github.com/Adleof/CVPM-2023-Temporal-Encoder-Decoder-iPPG.

1. Introduction

Physiological signals have been widely used in assessing vital metrics. Wearable devices have become increasingly popular because of their ease of use and the ability to provide continuous monitoring of physiological signals [20]. In recent years, there has been a surge of interest in developing non-intrusive methods for remotely monitoring vital signs and assessing physiological signals. The ongoing COVID-19 pandemic has further highlighted the importance of reducing physical contact while monitoring vital signs [36]. Recent advances in deep learning techniques have made it possible to extract vital signs from camera-based systems with relatively high accuracy [21]. Unlike traditional methods for monitoring physiological signals, camera-based systems have the potential to be less intrusive in many situations [29]. They can capture physiological signals using a standard camera or webcam, without the need for physical contact with the body. Cameras are therefore ideal for remote monitoring applications, as well as for situations where the use of other sensors is impractical or inconvenient.

Researchers have been actively working on developing imaging-based systems that can extract heart rate and other vital signs [7, 21, 32]. These efforts include photoplethysmography (PPG), in which light is used to measure changes in volume. When measured from a distance, the process is called remote image-based photoplethysmography (iPPG). In this work, the volumetric changes of interest are blood volume pulses (BVP) that result from the pumping action of the heart.

While many existing systems focus on estimating average heart rate or other basic vital signs, a more challenging problem is the recovery of BVP signal morphology from a video of the human face. Such information contains the shape and features of individual pulses in the BVP signal and can be used for a range of applications, such as biometric authentication [26], emotion detection [6], and stress monitoring [8]. However, recovering BVP signal morphology from video data is a difficult problem that requires careful temporal processing and signal analysis.

Bridging this gap, we propose a new approach that helps in getting shape morphology through information extraction from selected regions of interest (ROI) along with robust noise elimination. Whereas most previous methods have analyzed the entire face region or some large rectangular regions on the face, we have used an approach based on 3D face landmark estimation. We have also adapted a temporal

encoder-decoder model from the audio source separation domain to address the noise reduction task. Our results indicate PPG estimation with much higher fidelity than previous methods. We also propose a shape morphology metric that can be used for comparing PPG signals. Applications such as biometric authentication can benefit from such a metric.

To summarize, our main contributions are the following:

- This work is the first to apply a 1-dimensional temporal encoder-decoder model to the problem of iPPG signal extraction, and we demonstrate the ability to extract pulse shapes using this approach from videos of the face.
- We propose a skin reflection model that accounts for positioning noise introduced by the face landmark estimation procedure.
- We demonstrate that the extracted pulse shapes contain enough detail to support biometric authentication.

2. Background and Related Work

In recent years, there has been increasing interest in extracting and using PPG signals. Due to their easy acquisition and rich information about the cardiovascular system, PPG signals have been explored for such tasks as user authentication [13]. Unlike traditional biometric modalities such as fingerprints or face recognition, PPG signals are not visible to the unaided eye and are relatively difficult to steal or spoof [15]. PPG signals contain distinctive patterns of pulse waveforms generated by the heart's contractions and relaxations. The features specific to an individual's cardiovascular system, such as the shapes of diastolic and systolic peaks, can be used to create a biometric template for an individual that can be compared with templates from other individuals for biometric authentication [24–26].

Additionally, PPG signals can be obtained remotely using standard video cameras, allowing for convenient and non-invasive data collection [21]. Whereas traditional contact-based sensors provide PPG signals with more accuracy, some researchers have argued that remotely obtained PPG information is distinctive enough for biometric authentication [15, 26]. However, the problem is challenging because camera-based PPG signals are susceptible to noise from various sources that can affect the accuracy and reliability of the measurements [18]. Apart from the primary sources of noise due to lighting conditions, camera quality, motion artifacts, etc., noise also results from the facial vascular distribution and from differential neural controls by the sympathetic and parasympathetic nervous systems [10, 35]. As a result, there exist temporal differences in the cardiovascular activities in different parts of the face [16, 34]. The commonly used approach of averaging pulsating signals across large portions of the face, by considering the face as a single ROI, can thus potentially limit the ability to obtain distinctive pulse shapes [16]. We use multiple small ROIs on the face in an effort to address the heterogeneity of the facial vascular network.

PPG signal extraction from a camera is a challenging problem that must address unwanted noise that results from respiration, head movements, and variations in lighting conditions [1,31]. Thus, the problem can be posed as one of signal separation. For the separation of multiple signals in other domains, several techniques have been explored, such as Blind Source Separation [22, 23], Independent Component Analysis [9, 28], and non-negative matrix factorization [14, 33]. An example application is the separation of different sources in a music mixture [11]. Demucs, Deep Extractor for Music Sources, is a deep learning model that has been developed for music source separation tasks [5].

Demucs is a fully convolutional neural network that uses a technique called source-specific training to learn to separate the different sources in a music mixture. The architecture of Demucs consists of multiple layers of dilated convolutional neural networks, followed by a decoder network that uses transposed convolutional layers. The dilated convolutions allow the model to have a large receptive field while maintaining a small number of parameters, which is critical for the efficient processing of audio signals. Leveraging the effectiveness of Demucs in music source separation, a fully convolutional model based on Demucs could be trained on a dataset of video recordings to learn to separate the PPG signal from noise sources such as head movements or variations in lighting conditions. This paper demonstrates that the model can learn to identify and remove noise sources using the spatial and temporal features of the video signal, similar to how Demucs separates different musical instruments sources in an audio signal. We use our domain-adapted Demucs model to separate and filter out noise from the PPG signals. The extracted PPG signal contains less noise than has been demonstrated from previous camera-based PPG techniques.

3. Proposed Approach

The proposed approach presented in this paper can be categorized into four key steps: data conversion, data preprocessing, encoder-decoder-based extraction, and data postprocessing. The overview of our proposed approach is depicted in Figure 1.

3.1. Data Conversion

The first step, data conversion, involves the extraction of the average color intensities from Regions of Interest in every frame. The system extracts data from an image sequence, which is a 4-dimensional tensor with size $(num\_of\_frames,\ width,\ height,\ color\_channels)$ into a 2-dimensional color sequence tensor with size

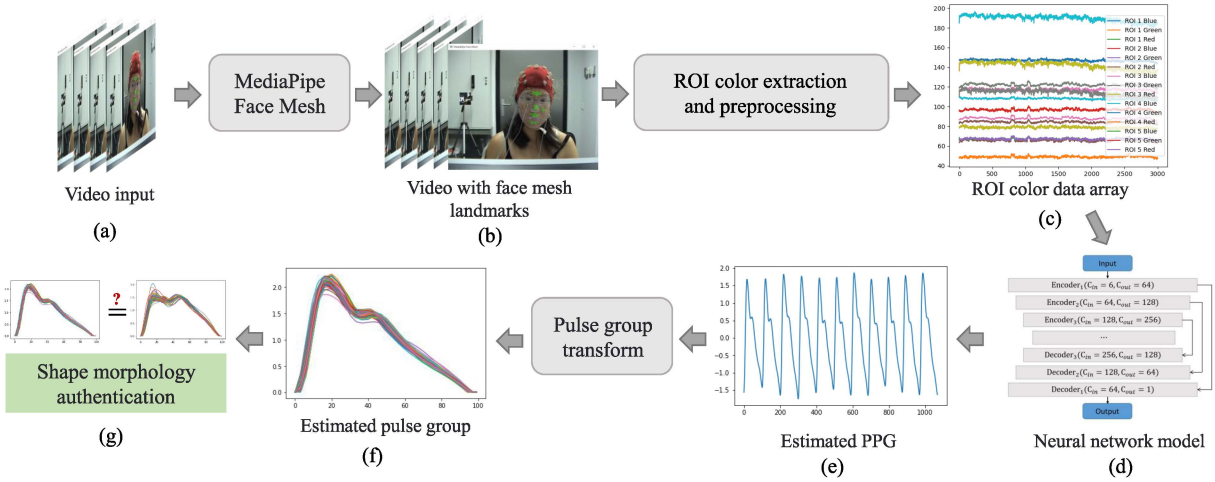


Figure 1. Our overall system structure. (a) The input to the system is a video. (b) ROIs are detected on individual frames, and (c) average values placed into an ROI color data array. (d) The neural network model takes the preprocessed data and generates (e) an estimated PPG signal. (f) Finally, in the postprocessing stage, the system extracts normalized PPG pulses from the estimated signal and (g) performs authentication using our shape morphology metric.

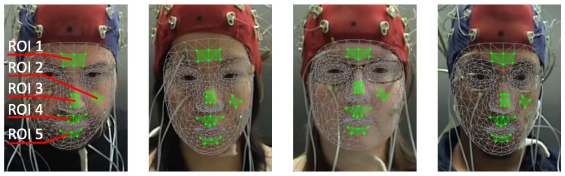


Figure 2. ROI visualization for different subjects. Five regions with high hemoglobin concentration were chosen empirically. These image frames are from the DEAP dataset [12].

$(num\_of\_frames, \quad color\_channels \times num\_of\_ROIs)$. The purpose of this stage is to remove irrelevant information that may interfere with the analysis of the PPG signal, such as head movement, hair color, and facial features. By extracting the average color of the ROIs, we obtain a color sequence that is more representative of the underlying PPG signal, enabling more accurate analysis in subsequent steps.

The DEAP dataset [12], used in our experiments, contains occlusions resulting from various confounding factors such as sensors and glasses. In order to select fixed non-occluded ROIs over time, we decided to generate meshes that are anatomically anchored and responsive to movement. The approach that we choose for ROI selection and data conversion is the MediaPipe Face Mesh [19] model by Google. This system fits a 3D triangular mesh to an image of the face. The MediaPipe Face Mesh outputs the pixel location and the depth of every vertex on the face UV map (we only use the pixel location in this work). By manually selecting particular vertices in the image, we obtain polygonal ROIs in the image that track head movements. We have only used selected ROIs that are rich in hemoglobin [34] and are not occluded by the sensors used for data collection. Examples of the ROIs on different subjects are shown in Figure 2.

3.2. Algorithm Design

This section describes our skin reflection model, preprocessing approaches, our temporal encoder-decoder model, and details of the experimental setup.

3.2.1 Skin Reflection Model

To evaluate the feasibility of iPPG using our approach, we need to model the video generation process. In order to account for the noise due to face mesh and model the noise elimination process with the encoder-decoder approach, the Skin Reflection Model from DeepPhys [2] was adapted for our task. We define the RGB values at the k-th ROI as a time-varying function:

$$\boldsymbol{C}_k(t) = I(t) \cdot (\boldsymbol{v}_s(t) + \boldsymbol{v}_d(t)) + \boldsymbol{v}_{nq}(t) + \boldsymbol{v}_{nf}(t) \quad (1)$$

where $\boldsymbol{C}_k(t)$ denotes a vector of RGB values; $I(t)$ is the luminance intensity level; $I(t)$ is modulated by the specular reflection $\boldsymbol{v}_s(t)$ and the diffuse reflection $\boldsymbol{v}_d(t)$; $\boldsymbol{v}_{nq}(t)$ denotes a mixture of the quantization noise and the background noise of the camera sensor; $\boldsymbol{v}_{nf}(t)$ denotes the positioning noise due to imperfections in face-mesh generation. The components $\boldsymbol{v}_s(t)$ and $\boldsymbol{v}_d(t)$ can be decomposed into stationary and time-dependent parts:

$$\boldsymbol{v}_s(t) = \boldsymbol{u}_s \cdot (s_0 + \Phi(m(t), p(t))) \quad (2)$$

$$\boldsymbol{v}_d(t) = \boldsymbol{u}_d \cdot d_0 + \boldsymbol{u}_p \cdot p(t) \quad (3)$$

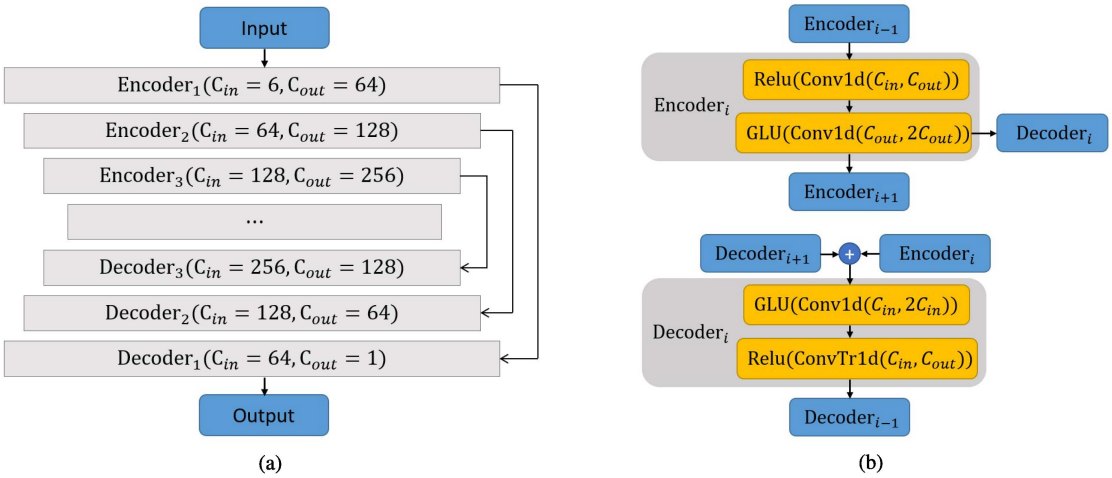


Figure 3. The network structure of our model, which was inspired by Demucs [5]. (a) Processing is performed by a U-Net structure that incorporates skip connections between encoder blocks and decoder blocks. Input to the system is a temporal sequence of (red, green, blue) values. The output is a temporal PPG waveform, with the same length as the input sequence. (b) Each encoder block consists of a 1D convolution layer with ReLU activation and another 1D convolution layer with gated linear unit (GLU) activation. The GLU activation will use the input data to modulate the output, which will introduce more non-linearity. The decoder block is the inverse of the encoder block. Inside each decoder block is a 1D convolution layer with GLU activation and a transposed 1D convolution with ReLU activation.

where $\boldsymbol{u}_s$ denotes the unit color vector of the light source spectrum; s_0 and $\Phi(m(t), p(t))$ denote the stationary and varying parts of specular reflections; $m(t)$ denotes all the non-physiological variations such as flickering of the light source, head rotation and facial expressions; and $p(t)$ denotes the PPG signal. Similarly, $\boldsymbol{u}_d$ denotes the unit color vector of the skin-tissue; d_0 denotes the stationary reflection strength; and $\boldsymbol{u}_p$ denotes the relative pulsatile strengths caused by hemoglobin and melanin absorption [2]. Substituting (2) and (3) into (1) we get:

$$\boldsymbol{C}_k(t) = I(t) \cdot (\boldsymbol{u}_s \cdot s_0 + \boldsymbol{u}_d \cdot d_0 + \boldsymbol{u}_s \cdot \Phi(m(t), p(t)) + \boldsymbol{u}_p \cdot p(t)) + \boldsymbol{v}_{nq}(t) + \boldsymbol{v}_{nf}(t) \quad (4)$$

In our authentication system we only consider stationary subjects with fixed lighting. Therefore $I(t)$ can be approximated by a constant I_0:

$$\boldsymbol{C}_k(t) \approx I_0 \cdot (\boldsymbol{u}_s \cdot s_0 + \boldsymbol{u}_d \cdot d_0) + I_0 \cdot \boldsymbol{u}_s \cdot \Phi(m(t), p(t)) + I_0 \cdot \boldsymbol{u}_p \cdot p(t) + \boldsymbol{v}_{nq}(t) + \boldsymbol{v}_{nf}(t) \quad (5)$$

This is our final equation for the skin reflection model. Our goal is to extract the PPG signal $p(t)$ from the ROI color sequences $\boldsymbol{C}_k(t)$. In (5), the first term $I_0 \cdot (\boldsymbol{u}_s \cdot s_0 + \boldsymbol{u}_d \cdot d_0)$ is a constant, which will be removed by band-pass filtering. The second term $I_0 \cdot \boldsymbol{u}_s \cdot \Phi(m(t), p(t))$ is non-linear, but this term can be assumed to be linear with respect to $p(t)$ when $m(t)$ is small [2], which is true under our assumptions of a stationary person with fixed illumination. The third term is also linear with respect to $p(t)$.

The last two terms are the quantization noise and the positioning noise. These signals are usually significant with respect to $p(t)$ and cannot be ignored. Therefore, after preprocessing, the task of the neural network model is to extract $p(t)$ from the preprocessed $\boldsymbol{C}_k(t)$, which is a combination of terms that have linear relationship with respect to $p(t)$ and the two noise terms. It is our hypothesis that the audio source separation model can extract the underlying PPG signal from this composite signal.

3.2.2 Data Preprocessing

We performed preprocessing of the ground-truth BVP signal and of the video sequences. The reason for preprocessing the ground-truth signals was large levels of noise from the BVP sensor, primarily low-frequency components that we removed through bandpass filtering. An additional preprocessing step normalized each pulse to have the same starting amplitude, ending amplitude, and same average amplitude.

Similarly, as discussed in section 3.2.1, bandpass filtering was applied to data that was obtained from the image sequence to remove a large constant term that is not related to the BVP signal.

3.2.3 Temporal Encoder-Decoder Model

The task of the temporal encoder-decoder neural network model stage is to extract the PPG signal from a mixture of the noise and terms that have a linear relationship with

respect to the PPG signal. As our base model, we developed a variation of Demucs [5], which is an audio source separation model. We believe that the audio separation task is quite similar to our signal extraction task.

In our system, the input to the neural network model is a data array in $\mathbb{R}^{L\times 3N}$ in which N denotes the number of ROIs and L denotes the sequence length. This array is formed by stacking preprocessed data sequences at all ROIs on their color channels. The output is a vector in $\mathbb{R}^{L}$ which is the estimated PPG signal.

Figure 3a shows the overall structure of our model, which is a U-Net structure with skip connections between encoder blocks and decoder blocks. The detailed structure inside each encoder and decoder block is shown in Figure 3b. For each encoder block, the first 1D convolution layer with ReLU activation will mainly be used as a downsampling layer. The second 1D convolution layer with the Gated Linear Unit (GLU) [3] activation is the core part of data processing. The GLU is a special kind of activation function that will use the input data to modulate the activation. This can introduce more non-linearity, and it is very useful in temporal sequence processing. The decoder block is the inverse of the encoder block. The aggregated data will go through a 1D convolution layer with gated linear unit activation and then up-sampled using a transposed 1D convolution with ReLU activation.

3.2.4 Loss Function

We define our loss function for training as follows:

$$L(y,\hat{y}) = (1-\lambda)\|y-\hat{y}\| + \lambda(\|Re(Y)-Re(\hat{Y})\| + \|Im(Y)-Im(\hat{Y})\|) \quad (6)$$

where y and $\hat{y}$ denote the output and the ground truth, respectively; Y and $\hat{Y}$ denote the Fourier transforms of the output and the ground truth. $Re(\cdot)$ and $Im(\cdot)$ denote the real part and the imaginary part of the input sequence. This loss function is a mixture of the mean squared error (MSE) loss and the Fast Fourier Transform (FFT) loss. In the FFT loss, we take the Fourier transform of both the output and the ground truth, then back propagate the mean squared error of the real part and the imaginary part separately using the auto-gradient module in PyTorch. We add the FFT loss because there are some details in the PPG signal such as the diastolic peak that are not obvious in the time domain. Therefore, creating a loss function in the frequency domain helps the model learn the PPG shape morphology better.

3.3. Postprocessing

This section describes the method for quantitatively analyzing the shape morphology, along with steps for postprocessing the signals.

3.3.1 Shape (Morphology) Metric

The morphology metric is one of the most important parts of our authentication process. As mentioned in Section 2, cardiovascular signals from each person are unique, and this fact motivates us to use the average shape of pulses in the PPG signal as one's biometric signature.

The system will convert the raw PPG signals that have been estimated by the neural network model to a PPG "pulse group", and then use this pulse group to extract a representative identifier of the individual. Figure 1e shows an example PPG signal of a person. Figure 1f shows the corresponding pulse group representation of the same signal. It is important to note that each pulse in the original PPG signal might have a different duration, mean value, and ending amplitude. After transformation, the pulses are scaled to have the same mean value, and their duration, starting, and ending amplitude are also be interpolated to be the same.

We define the pulse group transformation function as $f_g(\cdot)$. The input is the PPG signal $X \in \mathbb{R}^N$, where N denotes the sequence length. After applying transformation $f_g(X)$, we get the transformed signal group $X_g \in \mathbb{R}^{K\times L}$, where K denotes the total number of angular positions, and L denotes the number of pulses in the input signal. After the pulse group transformation, we further extract a person's identifier from the pulse group. In our system, we use the mean and the standard deviation at each angular position of the pulse group as the ID. We define $X_{Id} \in \mathbb{R}^{K\times 2}$, which refers to 2 numbers in each angular entry of X_{Id} representing the mean and the standard deviation of the same angular entry in X_g.

With the pulse group representation of the PPG signal, we can define the distance between a single pulse with a pulse group. Let X_g be the PPG signal group transformed from our estimation model and Y_{Id} be the ID of the target person we want to authenticate. Let $x_{gi} \in \mathbb{R}^K$ be a single normalized pulse in the pulse group X_g. We define the distance:

$$d(x_{gi}, Y_{Id}) = \frac{1}{K-2}\sum_{j=1}^{K-1}\frac{|x_{gij} - Y_{Idj}[mean]|}{Y_{Idj}[var]} \quad (7)$$

The symbol K denotes the total number of angular positions, $Y_{Idj}[mean]$ and $Y_{Idj}[var]$ indicate the mean and variance on angular entry j, respectively. Using this distance we measure how many standard deviations is our new pulses away from the reference signal at every angular position. Note that the summation excludes the first and the last angular positions, because the starting and ending amplitude are interpolated to be zero for all pulses, the variance on these two points will always be zero. Figure 4 shows two example distance measurements. The graph on the left shows an example of a new pulse that closely matches the

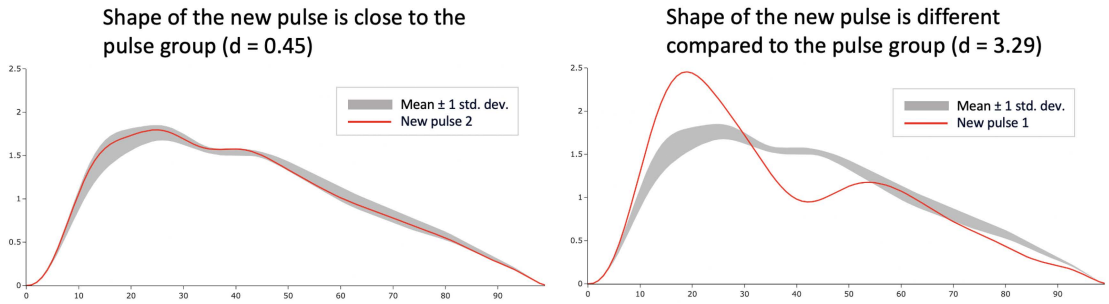


Figure 4. Example of the pulse distance measure. The left figure shows the case when a new pulse has a shape that is very similar to the reference pulse group. The right figure shows a case for which a new pulse has a very different shape compared to the reference pulse group.

reference pulse group Y_{Id}. The graph on the right shows an example of a new pulse with a very different shape compared to the reference pulse group. The distance calculated using our definition is 0.45 for the left pulse and 3.29 for the right pulse.

With this definition of distance between a single pulse and a pulse group, there are several ways to define group-to-group distances. For example, we can simply define the group-to-group distance as the mean of all pulse-to-group distances of pulses in X_g with Y_{Id}. An alternate way of defining group-to-group distance can be using the percentage of pulses-to-group distances that is higher than a threshold distance.

3.3.2 Noise Rejection

For every single pulse X_{gi} in X_g, we can calculate the pulse-to-group distance $d(X_{gi}, X_g)$ and remove pulses with a distance larger than some threshold. Note that removing pulses from a group can affect the mean and variance of X_g, therefore we choose to remove just one pulse in every iteration and iterate until no pulse can be removed.

4. Experiments

In this section, the experimental settings for our paper along with the results are explained. We have done a detailed analysis of shape recovery and the model's effectiveness in heart rate estimation.

4.1. Experimental Setting

We use the DEAP [12] dataset for training and testing. We chose DEAP for several reasons:

- Videos in this dataset have a high frame rate (50 fps) and a low level of compression.
- Participants did not exhibit significant body movements, and there are no significant lighting changes within the videos.
- The dataset provides an amount of video and ground truth PPG signals that is adequate for large-scale DNN training.

The entire dataset consists of 22 different participants with 874 one-minute videos. Each participant has 39-40 videos. We divided our data into training and testing sets with 714 videos from 18 participants for training, and the rest of the 160 videos from 4 participants for testing.

As described previously, we preprocess ground-truth PPG signals using bandpass filtering and pulse height normalization. The typical frequency band for the PPG signal is above 0.75 Hz (45 BPM), and so we use a high pass filter with a 0.75 Hz cut-off frequency. We manually checked the filtered signals to make sure that they retains the PPG pulse shape. Finally, the filtered signals are resampled to fit the sample rate of the video, which in our case is 50 Hz.

During training, the augmentation method that we used was random temporal clipping, so that each time the model saw different parts from the same data sequence. We adapted the Demucs model to have 15 channels (3 channels for each of the 5 ROIs) as input and 1 channel as output. We used 4 encoder/decoder layers. Each Conv1D/ConvTr1D layer in the encoder/decoder block has kernel size 10 and stride 2 to cover the entire PPG pulse range. We used the Adam optimizer with $lr = 0.0003$. We trained the model for 400 epochs using MSE loss between the estimated signal and ground truth.

4.2. PPG Shape Recovery

In previous rPPG methods, the raw output signal needs to be processed using bandpass filtering and frequency domain transformation. But in our method, the raw output from the neural network model already has a clear visible PPG shape morphology and can be transformed into a PPG pulse group directly. As an example, Figure 5 shows a comparison of the raw estimated signal morphology from our system and from MTTS-CAN [17]. In our output, the starting and the ending

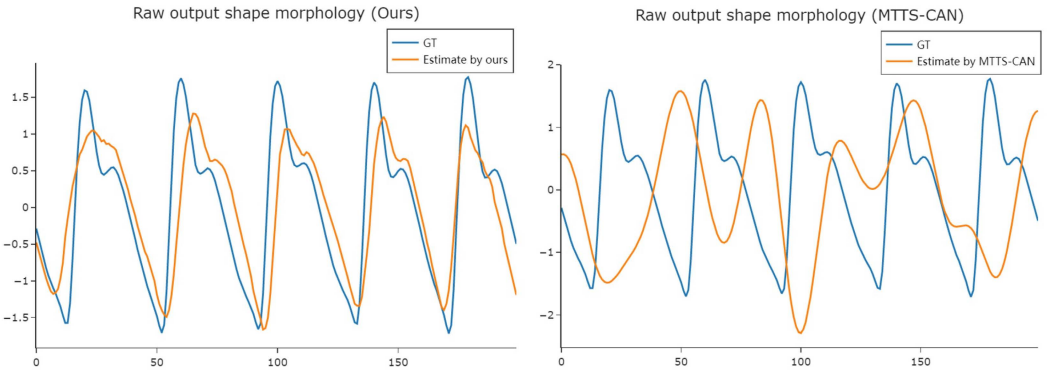
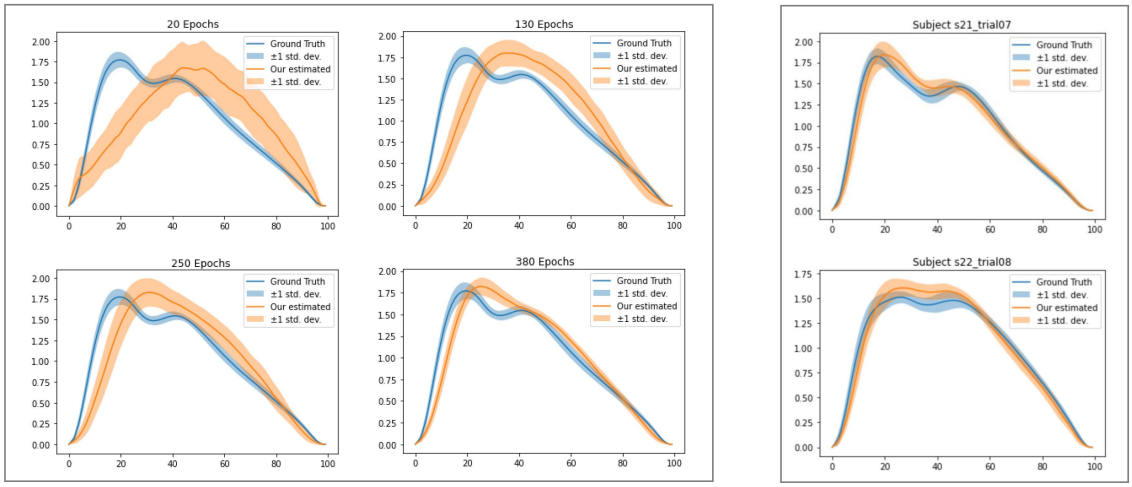


Figure 5. Example PPG estimations for the same subject using different methods. On the left is our result, and on the right is the result from MTTS-CAN. Our results show clear PPG pulses that could be used directly in applications such as instantaneous HR estimation.



(a) (b)

Figure 6. (a) The transformed PPG pulse group for the same test subject but with different amounts of training. With more and more training, our model gradually learns the correct PPG shape with lower variance. (b) Two transformed PPG pulse groups for 2 different subjects from the testing set.

points of each pulse are clear, while in many other previous methods the separate PPG pulses are not easily identified.

The ability to output clear PPG signals facilitates several possible research directions. One of them is biometric authentication using PPG. Previous research had found that the normalized PPG pulse shape from each person is unique [26], which implies that PPG may be a good modality for biometric authentication. Our method allows us to estimate the normalized pulse shape by applying the pulse group transform on the raw output directly. Figure 6 shows examples of estimated normalized PPG shapes from our model. Figure 6a shows the shapes for the same subject with increasing numbers of training epochs. With more and more training, the model gradually learns the correct PPG shape with lower variance. Similarly, Figure 6b shows two transformed PPG pulse groups for 2 different subjects from the testing set. The result demonstrates that the model is able to learn distinct shape morphology from different subjects.

4.3. Authentication

Using the recovered normalized pulse shapes from the model, we explored the feasibility of biometric authentication based on BVP shape alone. Figure 7 shows the ROC curve of the result. In this experiment, the authentication target is one particular subject from the test set using the PPG sensor data. All other subjects from the test set are compared with the target using their estimated pulse group from the model. The similarity score is calculated using the pulses in the pulse group that have a shape distance smaller than $d_{threshold} = 1.3$. The AUC value for this curve was approximately 0.77. Although this result is not

adequate for practical authentication tasks, the work clearly shows the potential for using rPPG-based methods as a new authentication method. It is our belief that some amount of overfitting to the dataset has limited the authentication performance, and additional training with a larger dataset will improve these results.

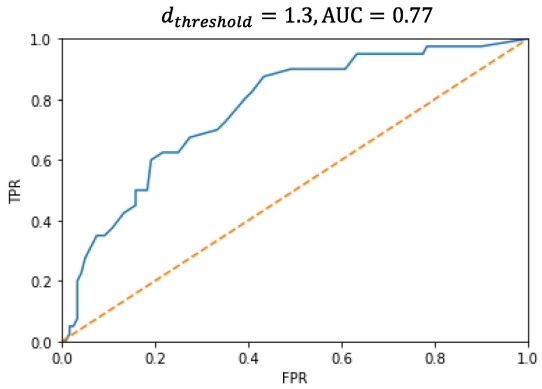


Figure 7. Using rPPG for biometric authentication. The authentication target is s19_trial06 [12] from the DEAP test set using PPG sensor data. The similarity scores were calculated using the pulses in the pulse group that have a shape distance smaller than the threshold distance $d_{threshold} = 1.3$. Distance was calculated using (7).

4.4. Estimating Heart Rate

The main objective of this paper has been to describe a new method for extracting PPG signals; however, our model's effectiveness has also been evaluated for the estimation of heart rate. We have compared our model's result to other systems using traditional heart-rate metrics using units of beats per minute. Detailed results of our technique with two ROIs (ROI 1 and ROI 2 from Figure 2) and five ROIs along with a comparison to state-of-the-art models are shown in Table 1. In our evaluation, a first-order Butterworth filter with cut-off frequencies of 0.7 and 2.5 Hz was applied to the model outputs. Then the filtered signals were divided into 5-second windows, and three standard metrics were computed over all windows of all the test videos in the dataset: mean absolute error (MAE), root mean square error (RMSE), and signal-to-noise ratio (SNR) of the estimated physiological signals averaged among all windows. The SNR is calculated in the frequency domain as the ratio between the energy for 0.2 Hz frequency bins around the gold-standard heart rate, and 0.05 Hz frequency bins around the gold-standard breathing rate [2].

The comparative results of our model with other models in terms of determining heart rate show that our model is also capable of estimating heart rate better than existing state-of-the-art models. Our model was trained on a unique loss function, which is a combination of MSE loss and FFT loss. This compound loss function helped the model to learn both heart rate estimation as well as shape morphology.

Method	$MAE_{bpm}\downarrow$	$RMSE_{bpm}\downarrow$	SNR (dB)↑
GREEN [30]	11.22	13.89	-8.07
CHROM [4]	9.70	12.45	-6.04
POS [31]	12.92	16.14	-9.46
HR-CNN [27]	15.91	18.75	-10.38
MTTS-CAN [17]	11.52	14.22	-7.65
Ours (2-ROI)	14.51	17.48	-9.99
Ours (5-ROI)	**9.41**	**11.26**	**-5.36**

Table 1. Heart rate recovery results using the DEAP dataset. The bottom rows indicate a significant performance improvement for our system when using 5 ROIs as input, compared with initial attempts using only 2 ROIs. (MAE = Mean Absolute Error, MSE = Mean Square Error, SNR = Signal-To-Noise Ratio)

4.5. Dataset Limitations

Currently, our model is capable of extracting clean PPG signals with good shape morphology using the DEAP dataset. Our proposed system can easily be adjusted to accept more ROIs and potentially achieve better performance. However, many potential ROIs in the DEAP dataset are occluded due to sensors worn on the face. We are currently working to create a novel dataset with fewer occlusions of the face [1]. The new dataset should be of value in further research that involves camera-based PPG sensing.

5. Conclusion

This paper has introduced a new approach for PPG signal shape extraction from RGB videos of a person's face. The system is the first to utilize a 1-dimensional temporal encoder-decoder model for remote (camera-based) PPG sensing. We also proposed a method to quantitatively measure the pulse morphology similarity between PPG signals.

Our experiments demonstrated good potential for using recovered PPG signals for the task of biometric authentication. In addition, the recovered PPG signals allowed us to estimate heart rate with better accuracy than previous state-of-the-art systems. Overall, our model's ability to estimate heart rate and to recover signal morphology makes it a promising tool for both health monitoring and authentication tasks.

Acknowledgement

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

[1] https://sites.google.com/view/vt-tricam-ppg

References

[1] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro d'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. An open framework for remote-PPG methods and their assessment. *IEEE Access*, 8:216083–216103, 2020. 2

[2] 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. 3, 4, 8

[3] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language Modeling with Gated Convolutional Networks. *CoRR*, abs/1612.08083, 2016. 5

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

[5] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019. 2, 4, 5

[6] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. 1

[7] Fridolin Haugg, Mohamed Elgendi, and Carlo Menon. GRGB rPPG: An Efficient Low-Complexity Remote Photoplethysmography-Based Algorithm for Heart Rate Estimation. *Bioengineering*, 10(2):243, 2023. 1

[8] Talha Iqbal, Andrew J. Simpkin, Davood Roshan, Nicola Glynn, John Killilea, Jane Walsh, Gerard Molloy, Sandra Ganly, Hannah Ryman, Eileen Coen, Adnan Elahi, William Wijns, and Atif Shahzad. Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset. *Sensors*, 22(21):8135, 2022. 1

[9] Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological measurement*, 26(1):R15, 2004. 2

[10] Hideaki Kashima, Tsukasa Ikemura, and Naoyuki Hayashi. Regional differences in facial skin blood flow responses to the cold pressor and static handgrip tests. *European Journal of Applied Physiology*, 113:1035–1041, 2013. 2

[11] Daichi Kitamura, Hiroshi Saruwatari, Kosuke Yagi, Kiyohiro Shikano, Yu Takahashi, and Kazunobu Kondo. Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 97(5):1113–1118, 2014. 2

[12] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 3, 6, 8

[13] Ruggero Donida Labati, Vincenzo Piuri, Francesco Rundo, and Fabio Scotti. Photoplethysmographic biometrics: A comprehensive survey. *Pattern Recognition Letters*, 2022. 2

[14] Daniel Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. *Advances in neural information processing systems*, 13:556–562, 2001. 2

[15] Lin Li, Chao Chen, Lei Pan, Jun Zhang, and Yang Xiang. SoK: An Overview of PPG's Application in Authentication. *CoRR*, abs/2201.11291, 2022. 2

[16] Jiangang Liu, Hong Luo, Paul Pu Zheng, Si Jia Wu, and Kang Lee. Transdermal optical imaging revealed different spatiotemporal patterns of facial cardiovascular activities. *Scientific Reports*, 8(1):1–10, 2018. 2

[17] 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. 6, 8

[18] Hao Lu, Hu Han, and S. Kevin Zhou. Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12404–12413, June 2021. 2

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines. *CoRR*, abs/1906.08172, 2019. 3

[20] Zhihan Lv and Yuxi Li. Wearable sensors for vital signs measurement: a survey. *Journal of Sensor and Actuator Networks*, 11(1):19, 2022. 1

[21] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):1–40, 2023. 1, 2

[22] Ganesh R. Naik, Wenwu Wang, et al. Blind source separation. *Berlin: Springer*, 10:978–3, 2014. 2

[23] Madhab Pal, Rajib Roy, Joyanta Basu, and Milton S. Bepari. Blind source separation: A review and analysis. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–5. IEEE, 2013. 2

[24] Jorge Sancho, Álvaro Alesanco, and José García. Biometric authentication using the PPG: A long-term feasibility study. *Sensors*, 18(5):1525, 2018. 2

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

[26] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Biometric authentication using photoplethysmography signals. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016. 1, 2, 7

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

[28] James V Stone. Independent component analysis: an introduction. *Trends in cognitive sciences*, 6(2):59–64, 2002. 2

[29] Rik van Esch, Kambez Ebrahimkheil, Iris Cramer, Wenjin Wang, Tomas Kaandorp, Federica Sammali, Angélique Dierick, 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

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

[31] 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, 2017. 2, 8

[32] Wenjin Wang, Steffen Leonhardt, Lionel Tarassenko, Caifeng Shan, and Daniel McDuff. Guest editorial: Camera-based monitoring for pervasive healthcare informatics. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1358–1360, 2021. 1

[33] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2012. 2

[34] Jing Wei, Hong Luo, Si J. Wu, Paul P. Zheng, Genyue Fu, and Kang Lee. Transdermal Optical Imaging Reveal Basal Stress via Heart Rate Variability Analysis: A Novel Methodology Comparable to Electrocardiography. *Frontiers in Psychology*, 9:98, 2018. 2, 3

[35] Thomas P. Whetzel and Stephen J. Mathes. Arterial Anatomy of the Face: An Analysis of Vascular Territories and Perforating Cutaneous Vessels. *Plastic and Reconstructive Surgery*, 89(4):591–603, 1992. 2

[36] Fan Yang, Shan He, Siddharth Sadanand, Aroon Yusuf, and Miodrag Bolic. Contactless Measurement of Vital Signs Using Thermal and RGB Cameras: A Study of COVID 19-Related Health Monitoring. *Sensors*, 22(2):627, 2022. 1