Abstract

We present frequency tracking for extracting heart rate trace from blood volume pulse (BVP) signal that can be used as an alternative for commonly used approach based on the mode of the BVP signal power spectral density. Our approach is based on particle filtering framework which provides smooth heart rate estimate, it is robust to motion-induced artifacts and noise. The method could be easily tuned and can be coupled with unsupervised BVP extraction approaches without the need for training. We evaluate our method on publicly available part of LGI dataset. Proposed algorithm shows competitive results comparing to argmax approach.

1. Introduction

Remote photoplethysmography (rPPG) is a technique for assessing blood volume changes in tissues by measuring small variations in reflected light that are detected with camera sensor. The rPPG signal itself can be useful for measuring vital signs like heart rate variability [12] or blood pressure [34] and, alongside with other bio-signals, can help in detecting various diseases [40]. In this paper we focus on heart rate estimation system pipeline.

Quality of PPG signal extracted from video depends on many factors like lighting conditions, camera parameters [24], body temperature, skin type and thickness of various skin tissue layers, subject movements and mimics [9]. The number and variability of that factors make it harder to gather representative dataset that takes them all into account.

Both deep learning based and traditional pipelines for assessing heart rate by video consist of multiple blocks, though the former tend to accumulate some of them in a single neural network which requires careful architectural as well as training procedure design. Performance of such multi-component system depends on the quality of each block in the pipeline, so it is important to analyze how separate component affects the results. Most of the recent papers devoted to blood volume pulse (BVP) prediction block [6, 7, 11, 22, 37–39, 41, 42], but some of them cover other components as well. To name a few, in [43] the impact of face detector and tracking on performance of heart rate prediction is analyzed; face landmark detection and semantic segmentation have been investigated in [30]; [3], among other improvements, includes analysis of filtering and power spectral density blocks, [13] provides an analysis of different facial ROIs; more detailed stratification of advances in components of heart rate prediction system can be found in [35]. In this paper we investigate the effect of frequency tracking on performance of heart rate prediction
pipelines.

The remainder of the paper is organized as follows. In Sec. 2, we provide an overview of the existing methods for remote video heart rate estimation and frequency tracking. Section 3 provides detailed description of the proposed solution. Section 4 describes implementation details. In Sec. 5, we present results of evaluation of the pipeline. Section 6 contains concluding remarks.

2. Related work

From the perspective of our research, solution pipelines for the task of remote video heart rate estimation can roughly be separated into two parts: the first one predicts BVP signal based on a sequence of face images; the second extracts heart rate frequency directly from the predicted BVP signal.

2.1. BVP signal prediction

A lot of methods for BVP signal prediction were proposed recently. Some classical, non deep learning based approaches try to estimate BVP by use of various matrix decomposition techniques like PCA [18] or ICA [31]. Recently Casado and Lopez proposed QR factorization based method called OMIT [3]. The derivation of another group of classical methods is based on dichromatic reflectance model, so they are called model-based. PBV [7], POS [39] and CHROM [6] are among the well-known representatives of this group. In POS intensity variations are firstly cancelled out and then pulsatile component is extracted by combining channel signals, while CHROM eliminates specular component assuming standardized skin-tone vector is known. PBV, instead, uses known blood volume pulse color direction to extract pulsatile component which is found by least squares fit. DIS [38] adds motion information to the signal matrix and searches for the projection vector which jointly minimizes motion-induced artifacts and maximizes pulsatile signal strength.

With a raising success of deep learning, neural network based solutions for BVP prediction gained a lot of attention. Physnet [41] proposed a family of rPPG prediction networks which includes separable spatio-temporal (2DCNN), 3D convolutions (3DCNN) and recurrent modules (LSTM, BiLSTM, ConvLSTM). DeepPhys [4] introduced 2D convolutional attention network (CAN) architecture which has two branches: motion branch predicts BVP signal, while relevant facial areas are extracted with appearance branch and injected to the motion branch by attention modules. Later, in [22] the family of CAN networks has been extended with 3D-CAN, Hybrid 2D-/3D-CAN and, eventually, TS-CAN, which exploits temporal shift modules instead of 3D convolutions in order to reduce computational burden. Physformer [42] utilizes transformer architecture that is able to catch much longer spatio-temporal interac-

2.2. Heart rate extraction

Since all the methods mentioned above in the section are targeted to predict BVP signal some additional post processing step is required to get the heart rate. Straightforward way would be to transform the signal to frequency domain with short-time Fourier transform (STFT) and analyze the spectrogram. Spectral peak or argmax estimator, is the most popular approach for heart rate extraction [3, 4, 6, 7, 11, 13, 22, 38, 42]. Sun and Li [37] additionally check spectrum for the presence of second harmonic and correct final prediction if needed. Another approach to compute heart rate is based on the inverse of average inter-beat interval of BVP signal [12, 18, 31, 41], but this method is more demanding to the quality of the signal.

Hsu et al. [15] take spectral representation of unsupervised signal, concatenate spectral features from neighbourhood windows and train SVR for heart rate prediction. In [14] 512-point BVP signal was transformed to frequency domain with STFT, spectrogram image was rescaled to the size of 128x128 and fed to 15-layered VGG network which predicts one of 200 classes of feasible heart rate values. Zhu et al. [45] design two-step procedure for heart rate prediction from spectrogram image. They first binarize spectrogram per 95th percentile at each moment, use morphological operations to connect broken traces and select the largest connected component as the most probable frequency strap. Next, they compute weighted average frequency, where the weight of each component is proportional to its relative power.

Spetlik et al. [36] proposed CNN-based heart rate estimator operating in temporal domain. The BVP extraction net has been fine-tuned for each benchmark dataset separately. Comparing
Various frequency tracking methods were developed for the task of speech and music signal analysis. Dubois et al. [8] use jump Markov system to model arbitrary number of frequency components in acoustic signal and STFT-based observation model in particle filtering framework. Fujimoto et al. [10] have developed pitch and harmonic frequencies tracking solution based on particle filter. Ng et al. [26] analyzed the effect of different dynamic models on the quality of single-tone frequency prediction. Kim et al. [17] use sigma-point Kalman smoother for multi-harmonic frequency tracking. Recently Das et al. [5] proposed Extended Kalman Filter with complex-valued state vector for monophonic pitch tracking. In [16] authors designed convolutional tracker that operates on time-domain waveform.

Numerous solutions in other fields have also benefited from frequency tracking. Nagappa and Hopgood [25] proposed single-tone frequency tracker for bat echolocation signal analysis. In [21] Rao-Blackwellized particle filtering approach was applied for the tasks of wheel vibration estimation and car engine sound frequency tracking. Sandberg et al. [33] developed HMM-based tracking algorithm for atrial fibrillation diagnostics. Zhu et al. [44] proposed dynamic programming based approach called Adaptive Multi-Trace Carving and validated the method on electric network and rPPG signal frequency estimation tasks. To the best of our knowledge, this is the only work which applies frequency tracking to the task of heart rate estimation. Different from the latter, our research has several distinctions: (i) our method is based on particle filtering algorithm from [25] and targeted to track only single frequency, (ii) the tracker in [44] needs the full frequency representation of BVP signal in range of interest, our method, instead, requires to compute only one coefficient corresponding to the current frequency estimate, (iii) they validate on simulated and private real-world fitness exercise dataset, where the BVP is extracted with CHROM [6] and do not test any other methods.

3. Proposed Approach

3.1. Particle Filter for heart rate tracking

Particle filtering is a well-known approach for estimating the state of dynamical system that can cope with non-linear dependencies in the models and non-Gaussian noise. For the task of heart rate estimation, such system tries to predict heart rate (state variable) based on noisy rPPG measurements (observations) derived from video. We use simple random-walk model for process dynamics:

$$f_t = f_{t-1} + v_{t-1},$$

Table 1. Average performance of particle filter(pf) and welch argmax(wa) estimators on LGI dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE_pf</th>
<th>MAE_wa</th>
<th>PCC_pf</th>
<th>PCC_wa</th>
<th>RMSE_pf</th>
<th>RMSE_wa</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICA</td>
<td>15.22</td>
<td>18.43</td>
<td>0.07</td>
<td>0.06</td>
<td>17.26</td>
<td>29.63</td>
<td>-0.22</td>
</tr>
<tr>
<td>PBV</td>
<td>14.93</td>
<td>14.27</td>
<td>0.26</td>
<td>0.21</td>
<td>16.66</td>
<td>22.55</td>
<td>0.21</td>
</tr>
<tr>
<td>PCA</td>
<td>12.30</td>
<td>10.99</td>
<td>0.24</td>
<td>0.38</td>
<td>14.72</td>
<td>15.46</td>
<td>2.13</td>
</tr>
<tr>
<td>CHROM</td>
<td>11.81</td>
<td>10.34</td>
<td>0.28</td>
<td>0.35</td>
<td>13.81</td>
<td>15.86</td>
<td>2.61</td>
</tr>
<tr>
<td>OMIT</td>
<td>11.31</td>
<td>8.98</td>
<td>0.32</td>
<td>0.38</td>
<td>13.03</td>
<td>14.10</td>
<td>3.05</td>
</tr>
<tr>
<td>LGI</td>
<td>5.39</td>
<td>9.02</td>
<td>0.44</td>
<td>0.39</td>
<td>7.49</td>
<td>14.13</td>
<td>3.06</td>
</tr>
<tr>
<td>POS</td>
<td>4.03</td>
<td>5.26</td>
<td>0.47</td>
<td>0.46</td>
<td>5.79</td>
<td>9.43</td>
<td>4.54</td>
</tr>
</tbody>
</table>
where \( f_t \) is a heart rate frequency at the moment \( t \), \( v_{t-1} \) is an additive noise component at the moment \( t-1 \). The observation equation
\[
z_t = h(f_t) + u_t,
\]
where \( z_t \in \mathbb{R}^n \) is a BVP signal window of \( n \) samples at moment \( t \), \( h : \mathbb{R} \to \mathbb{R}^n \) is a non-linear function, \( u_t \sim \mathcal{N}(0, \sigma_u^2) \) is an additive noise component at the moment \( t \).

To simplify the notation, let introduce the following discrete cosine and sine signals of frequency \( f_s \) sampled with the video frame rate \( f_s \):
\[
c = \cos \left( 2\pi \frac{f_t}{f_s} k \right), \quad k \in \mathbb{N} \quad (3)
\]
\[
s = \sin \left( 2\pi \frac{f_t}{f_s} k \right), \quad k \in \mathbb{N}. \quad (4)
\]

Then we could denote the windows of the same length and location as \( z_t \), but which are taken from these cosine and sine curves as \( c_t \) and \( s_t \), respectively. Using this notation the likelihood can be determined by the formula [25]:
\[
p(z_t|f_t) \propto \sigma_u^{-n+2} \exp \left( -\frac{z_t^T z_t - 2C}{2\sigma_u^2} \right), \quad (5)
\]
where \( C \) is Schuster periodogram coefficient at the frequency \( f_t \) defined as
\[
C = \frac{(z_t^T c_t)^2 + (z_t^T s_t)^2}{n}. \quad (6)
\]

In particle filtering framework the distribution of state-space variable is modeled with weighted samples generated from some initial distribution and propagated according to Eq. (1). The weights are updated with the likelihood:
\[
w_t^i = w_{t-1}^i p(z_t|f_t), \quad (7)
\]
where \( w_t^i \) is the \( i \)-th sample weight at the moment \( t \). We normalize the weights to represent probability distribution.

In order to prevent samples degeneracy problem we resample the particles on the following condition:
\[
\frac{1}{\sum_{i=1}^m (w_t^i)^2} < N_{\text{thresh}}, \quad (8)
\]
where \( N_{\text{thresh}} \) is threshold on effective number of particles. The final estimate is weighted sum of the samples
\[
\hat{f}_t = \sum_{i=1}^m w_t^i f_t^i, \quad (9)
\]
where \( m \) is the number of particles.

### 3.2. Initialization

The straightforward way to initialize state of the particle filter would be the mode of power spectral density (PSD) obtained for the first window of BVP signal. This requires the person to be still and whole setup is not changing during this period of time, which does not hold true in some cases. For such videos we could fit Gaussian Mixture Model to normalized PSD of the first window BVP signal and use these statistics to initialize \( L \) particle filters. Such curve fitting based solution would be sensitive to proper initialization and argument bounds, so we instead propose the following algorithm for initialization.

<table>
<thead>
<tr>
<th>MAE_pf</th>
<th>MAE_wa</th>
<th>PCC_pf</th>
<th>PCC_wa</th>
<th>RMSE_pf</th>
<th>RMSE_wa</th>
<th>SNR</th>
<th># of predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>gym</td>
<td>8.54</td>
<td>19.62</td>
<td>0.74</td>
<td>0.38</td>
<td>12.49</td>
<td>29.95</td>
<td>-2.16</td>
</tr>
<tr>
<td>talk</td>
<td>8.85</td>
<td>11.08</td>
<td>0.17</td>
<td>0.16</td>
<td>10.99</td>
<td>16.59</td>
<td>-0.02</td>
</tr>
<tr>
<td>rotation</td>
<td>2.71</td>
<td>3.94</td>
<td>0.39</td>
<td>0.28</td>
<td>4.20</td>
<td>7.43</td>
<td>4.90</td>
</tr>
<tr>
<td>resting</td>
<td>1.46</td>
<td>1.44</td>
<td>0.47</td>
<td>0.72</td>
<td>2.26</td>
<td>2.55</td>
<td>9.53</td>
</tr>
<tr>
<td>average</td>
<td>5.39</td>
<td>9.02</td>
<td>0.44</td>
<td>0.39</td>
<td>7.49</td>
<td>14.13</td>
<td>3.06</td>
</tr>
</tbody>
</table>

Table 2. Performance of particle filter (pf) and welch argmax (wa) estimators per each type of activity in LGI dataset; BVP signal is obtained with LGI method

<table>
<thead>
<tr>
<th>MAE_pf</th>
<th>MAE_wa</th>
<th>PCC_pf</th>
<th>PCC_wa</th>
<th>RMSE_pf</th>
<th>RMSE_wa</th>
<th>SNR</th>
<th># of predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>gym</td>
<td>2.42</td>
<td>8.50</td>
<td>0.96</td>
<td>0.66</td>
<td>4.50</td>
<td>16.08</td>
<td>2.04</td>
</tr>
<tr>
<td>talk</td>
<td>9.80</td>
<td>7.36</td>
<td>0.02</td>
<td>0.12</td>
<td>12.89</td>
<td>12.00</td>
<td>0.61</td>
</tr>
<tr>
<td>rotation</td>
<td>2.31</td>
<td>3.32</td>
<td>0.41</td>
<td>0.36</td>
<td>3.46</td>
<td>6.44</td>
<td>5.87</td>
</tr>
<tr>
<td>resting</td>
<td>1.58</td>
<td>1.85</td>
<td>0.49</td>
<td>0.70</td>
<td>2.32</td>
<td>3.18</td>
<td>9.63</td>
</tr>
<tr>
<td>average</td>
<td>4.03</td>
<td>5.26</td>
<td>0.47</td>
<td>0.46</td>
<td>5.79</td>
<td>9.43</td>
<td>4.54</td>
</tr>
</tbody>
</table>

Table 3. Performance of particle filter (pf) and welch argmax (wa) estimators per each type of activity in LGI dataset; BVP signal is obtained with POS method
At first step, we find the peak power in the spectrum of the initial half-length window of BVP signal and set all the components that are less than 5% of that value to zero. Next, the peak finding algorithm is applied to the spectrum in order to extract $L$ local maxima that are subsequently used for initialization of particle filters. We track posterior trajectories for some period of time $T$ and compute cumulative relative power along each one. Finally, the filter with the maximum cumulative power survives while all others are dropped. If $T$ is equal to the length of the video we get offline mode, while fixing $T$ to some reasonable value can be used as a calibration step in online mode.

4. Experiments

4.1. Dataset

Publicly available part of LGI dataset [29] has been used for model evaluation, it consists of 24 videos of 6 participants performing 4 different types of activities. Each video has duration not less than one minute, for gym type of activity videos are about five times longer than for other sessions. Ground truth heart rate estimates were obtained with CMS50E pulseoximeter.

4.2. Implementation details

Our implementation is based on pyVHR framework [1, 2]. For face area and landmark detection we choose default setting that is based on Mediapipe face mesh [23]. Whole face area excluding eyes and mouth regions is used for extracting colored signal (which is also named as holistic in the framework). Various classical methods like POS or CHROM were used for BVP signal computation subsequently. Sixth order Butterworth bandpass filter with 0.65 to 4 Hz passband is used for post-processing. We follow [3] and traverse predicted and ground truth BVP signals with the same window (both are centered and have equal length, we do not pad the signal) in order not to introduce additional time shift between the signals.

We test the pipeline with the window length of 10 seconds. $N_{thresh}$ was set to 20 and $m$ to 100 particles. Pro-
cess noise was sampled from normal distribution \( \mathcal{N}(0, \sigma_v^2) \), where \( \sigma_v \) was set to 0.5 bpm; standard deviation of measurement noise \( \sigma_w \) was set to 1.0, these values were found to produce appropriate smooth heart rate trajectory on alex_gym video, we have not precisely tuned the parameters. First samples within half-length window is used for filter initialization, so power spectral density is computed on first 5 seconds of the signal and then peak detection algorithm is applied to find proposal frequencies. We limit only the horizontal distance between the peaks to be not less than 10 bpm. We scale each windowed signal \( z_t \) with Hanning window function coefficients in order to reduce spectral leakage and normalize it after to have constant power.

4.3. Evaluation

Recent challenges on remote heart rate estimation [19, 20, 32] encourage to use heart rate level metrics, but the solutions mostly try to reduce the error by developing BVP prediction block coupled with standard argmax estimator. Earlier benchmarks [19] popularized average heart rate prediction task (only one value for single video is being predicted). They usually randomly cut longer video into short segments that are considered independently. Without access to original longer videos, such protocol seems to be not so inspiring for the solutions that utilizes temporal correlation in their predictions. Later challenges proposed continuous, e.g. frame-level heart rate prediction metric [32]. We adopt continuous metrics from pyVHR framework, where predicted and ground truth heart rates are compared once each second of time. We propagate process and observation models of particle filter for each frame and downsample predictions later with the factor equal to fps, when compute the metrics.

In order to investigate capabilities of classical methods combined with particle filter we start with offline operating mode by setting \( T \) to the length of the video. Due to stochastic nature of particle filtering we repeat each trial 50 times varying only random seed used for particles propagation according to Eq. (1) and compute error statistics of this ensemble predictions for each video. We obtain worst metric values across the ensemble for each video and then average them over all the videos according to the next formula:

\[
\frac{1}{AP} \sum_{i=1}^{A} \sum_{j=1}^{P} \max_r (M(f^r_{ijt} - \hat{f}^r_{ijt})),
\]

where \( M \) is the metric we would like to be minimal (e.g. RMSE or MAE), \( A \) is the number of activities performed by \( P \) participants in the dataset, \( f^r_{ijt} \) is the heart rate prediction for \( j \)-th participant performing \( i \)-th type of activity at the moment \( t \) by \( r \)-th realization; for PCC we substitute \( \max \) with \( \min \) in the formula.

5. Results and Discussion

We compare proposed frequency tracker with argmax method on Welch’s spectrogram. Results for different BVP extraction methods are presented in Tab. 1. We get substantial improvements of RMSE for ICA, PBV, LGI and POS, while for PCA, CHROM and OMIT the RMSE difference for the two heart rate extraction approaches is not so drastic. MAE achieves lower values for ICA, LGI and POS only. In general, with increasing SNR, prediction error goes down for both welch argmax and particle filter.

Next, we analyze the errors with regard to each type of activity for the two BVP methods that have shown best accuracy on the dataset, that are LGI and POS. It can be noticed from Tab. 2 and Tab. 3 that the main improvement is achieved for gym session videos, while for talk we even sometimes get worse results than welch argmax.

To estimate stability of particle filter predictions we draw mean and standard deviation of RMSE across different realizations in Fig. 2. For POS (Fig. 2b) heart rate traces are relatively stable for rotation and resting scenarios, while for other types of activities the method shows increased variance, especially on those videos that have negative SNR as shown in Fig. 3b. We also consider extreme values of mean RMSE in each activity group separately, check the corresponding videos and find that (i) angelo_gym and fe-
lix_talk videos have errors in reference BVP signal, which is also noticed in [3]; (ii) felix_resting video contains temporally varying illumination pattern and (iii) cpi_talk have exaustive facial area intensity variations caused by camera autotuning algorithms during head movements. For LGI an increased error variance for alex_gym video (Fig. 2a) is caused by trace switching phenomenon as seen in Fig. 4a. Here overall low SNR facilitates the switching to pedal rotation frequency trace, POS instead provides BVP signal with dominating heart frequency and the cost of switching in terms of accumulated power loss is higher as seen in Fig. 4b.

The algorithm performance depends on the type of noise, not only SNR. Though we obtain the worst SNR levels of BVP signal during gym session (Fig. 3), our algorithm is able to predict the right trace, since the heart rate track is preserved on the spectrogram and is separated from pedal rotation trace. Instead, for talk scenarios heart rate trace is not pronounced and the parasitic traces are located closely to it. Though the errors of particle filter predictions for talk videos with POS are higher than the argmax method (Tab. 3), our predictions do not contain spurious values and overall trajectories are smooth.

Having presented the results in offline mode, we now analyze the effect of different values of power accumulation period $T$ on predictions. Since heart rate trace could be less pronounced for some period of time, reducing the $T$ results in error increase (Fig. 5). As expected, for resting type of activity there is no drop in accuracy, while for others the errors grow due to confusion of heart rate with temporally dominating parasitic traces preserved in BVP signal. In Fig. 6, it can be clearly seen how reduction of power accumulation period from 40 to 20 seconds results in switching to pedal rotation frequency trace. Motion frequency notching could potentially mitigate the effect in such cases.

Finally, if $T = 0$ we do not use power accumulation scheme and select only one particle filter initialized with argmax frequency on the first 5 seconds of the video. Still, for POS restricting $T$ to 40 seconds does not result in any noticeable deterioration comparing to the case when the trace is selected by power accumulation along the full length of the video.

6. Conclusion

In this paper we propose single-tone frequency tracking approach based on particle filtering framework for remote heart rate estimation. Our results demonstrate that the prediction accuracy can be substantially improved if we utilize temporal correlation between the neighboring heart rate estimates. Our initialization algorithm provides proper heart trace selection in challenging gym scenarios and results in
Figure 6. Effect of power accumulation threshold (shown as yellow dashed vertical line) on heart rate predictions for harun_gym video (black - particle filter realizations, red - ground truth). BVP signal is obtained with POS

reduced error variance.

Acknowledgements

The authors would like to thank Maxim Kazakov (Huawei) for valuable feedback and insights to accomplish this work.

References

In 2011 federated conference on computer science and information systems (FedCSIS), pages 405–410. IEEE, 2011. 2


[40] Bing-Fei Wu, Bing-Jhang Wu, Shao-En Cheng, Yu Sun, and Meng-Liang Chung. Motion-robust atrial fibrillation detection based on remote-photoplethysmography. IEEE Journal of Biomedical and Health Informatics, 2022. 1


[43] Changchen Zhao, Peijie Mei, Shoushuai Xu, Yongqiang Li, and Yuanjing Feng. Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography. In Proceedings of the IEEE/CVF Interna-