Frequency Tracker for Unsupervised Heart Rate Estimation

Iskander Zhalbekov[1], Leonid Beynenson[2], Alexey Trushkov[1], Ivan Bulychev[2], and Wenshuai Yin[1]
[1]Huawei CBG AI
{zhalbekov.iskander1,beynenson.leonid,trushkov.alexey,yinwenshuai}@huawei.com
bulychev.ivan@huawei-partners.com

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 photopletysmography (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

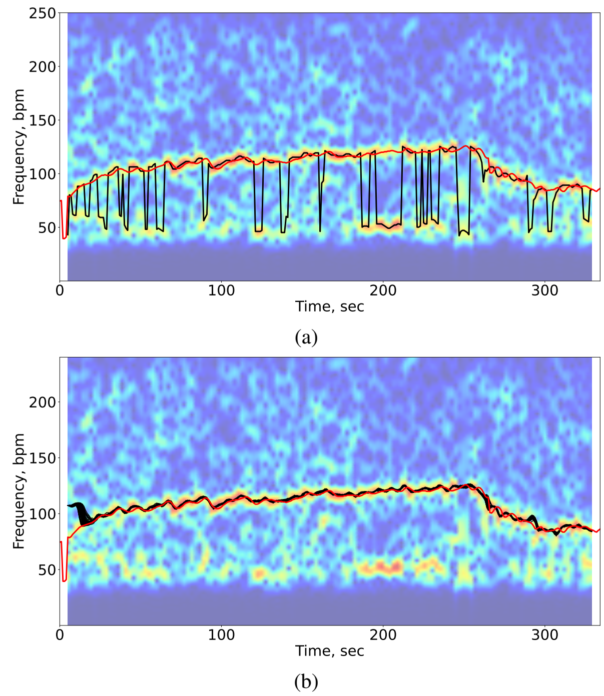


Figure 1. Heart rate prediction on david_gym video with different heart rate extraction approaches: (a) argmax estimator (black) (b) particle filter (50 realizations shown in black); spectrograms are obtained for BVP by LGI, ground truth is shown in red

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 interactions for BVP signal modeling but the resulting network is not suitable for operating on mobile device.

Lack of labeled data for training deep learning methods can be mitigated by use of self-supervised learning. Gideon and Stent [11] resample video with random factor that results in changing the perceived heart rate. Modified version of 3DCNN-based PhysNet was trained with supervised cross-correlation loss along with triplet loss where BVP predicted for temporally up-/downsampled videos constitute pool of negative samples and the BVP signals resampled back to original sampling frequency are used as positive examples. In contrast-phys [37] 3DCNN-based PhysNet was trained with contrastive loss only, making the model completely unsupervised. BVP signals for the same video but for different facial areas are being pulled together while signals corresponding to different videos are being repelled from each other. Such kind of training is sufficient to achieve the results comparable to supervised solutions on some rPPG benchmarks.

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 feeded 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 freezed while heart rate estimator was finetuned for each benchmark dataset separately. Comparing

	MAE_pf	MAE_wa	PCC_pf	PCC_wa	RMSE_pf	RMSE_wa	SNR
ICA	**15.22**	18.43	**0.07**	0.06	**17.26**	29.63	-0.22
PBV	14.93	**14.27**	**0.26**	0.21	**16.66**	22.55	0.21
PCA	12.30	**10.99**	0.24	**0.38**	**14.72**	15.46	2.13
CHROM	11.81	**10.34**	0.28	**0.35**	**13.81**	15.86	2.61
OMIT	11.31	**8.98**	0.32	**0.38**	**13.03**	14.10	3.05
LGI	**5.39**	9.02	**0.44**	0.39	**7.49**	14.13	3.06
POS	**4.03**	5.26	**0.47**	0.46	**5.79**	9.43	4.54

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

to argmax estimator authors claim CNN-based approach is more robust to non-stationary and noisy heart rate traces. They compare the network pulse predictions to some classical methods, but do not train CNN-based estimator with the output of that algorithms. RhythmNet [27] takes into account succeeding predictions by adding GRU-layer on top of the predicted convolutional features and imposing smoothness constraints on final heart rate trace. Niu *et al*. [28] proposed two separate convolutional heads for BVP and average heart rate prediction that exploit common intermediate feature map.

Since argmax estimator is very sensitive to presence of outliers, performance of the pipeline which uses such estimator depends on the quality of BVP signal. One way to mitigate the affect of temporary disturbance in BVP signal is to increase the width of sliding window in STFT, that would suppress spurious frequency components but lower temporal resolution in predictions. In more complex scenarios, when BVP signal contains parasitic frequency components, that, for instance, correspond to periodic movements during fitness training, argmax estimator could produce multiple abrupt unnatural jumps from one frequency track to another as shown in Fig. 1a. Movement harmonic cancellation might going to help but there are few studies investigating such cases.

Supervised methods could potentially provide smooth heart rate trace, but require training dataset. We instead propose classical signal processing algorithm which can be easily tuned and used together with unsupervised rPPG algorithm.

2.3. Frequency tracking

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 have 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, 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:

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

	MAE_pf	MAE_wa	PCC_pf	PCC_wa	RMSE_pf	RMSE_wa	SNR	# of predictions
gym	**8.54**	19.62	**0.74**	0.38	**12.49**	29.95	-2.16	1967
talk	**8.85**	11.08	**0.17**	0.16	**10.99**	16.59	-0.02	428
rotation	**2.71**	3.94	**0.39**	0.28	**4.20**	7.43	4.90	369
resting	1.46	**1.44**	0.47	**0.72**	**2.26**	2.55	9.53	362
average	**5.39**	9.02	**0.44**	0.39	**7.49**	14.13	3.06	

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

	MAE_pf	MAE_wa	PCC_pf	PCC_wa	RMSE_pf	RMSE_wa	SNR	# of predictions
gym	**2.42**	8.50	**0.96**	0.66	**4.50**	16.08	2.04	1967
talk	9.80	**7.36**	0.02	**0.12**	12.89	**12.00**	0.61	428
rotation	**2.31**	3.32	**0.41**	0.36	**3.46**	6.44	5.87	369
resting	**1.58**	1.85	0.49	**0.70**	**2.32**	3.18	9.63	362
average	**4.03**	5.26	**0.47**	0.46	**5.79**	9.43	4.54	

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

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

$$\mathbf{z}_t = \boldsymbol{h}(f_t) + \mathbf{u}_t, \tag{2}$$

where $\mathbf{z}_t \in \mathbb{R}^n$ is a BVP signal window of n samples at moment t, $\boldsymbol{h} : \mathbb{R} \to \mathbb{R}^n$ is a non-linear function, $\mathbf{u}_t \sim \mathcal{N}(0, \sigma_u^2)$ 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_t sampled with the video frame rate f_s:

$$c = \cos\left(2\pi \frac{f_t}{f_s} k\right), k \in \mathbb{N} \tag{3}$$

$$s = \sin\left(2\pi \frac{f_t}{f_s} k\right), k \in \mathbb{N}. \tag{4}$$

Then we could denote the windows of the same length and location as $\mathbf{z}_t$, but which are taken from these cosine and sine curves as $\mathbf{c}_t$ and $\mathbf{s}_t$ respectively. Using this notation the likelihood can be determined by the formula [25]:

$$p(\mathbf{z}_t|f_t) \propto \sigma_u^{-n+2} \exp\left(-\frac{\mathbf{z}_t^T \mathbf{z}_t - 2C}{2\sigma_u^2}\right), \tag{5}$$

where C is Schuster periodogram coefficient at the frequency f_t defined as

$$C = \frac{(\mathbf{z}_t^T \mathbf{c}_t)^2 + (\mathbf{z}_t^T \mathbf{s}_t)^2}{n}. \tag{6}$$

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:

$$w_t^i = w_{t-1}^i p(\mathbf{z}_t|f_t), \tag{7}$$

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:

$$\frac{1}{\sum_{i=1}^n (w_t^i)^2} < N_{thresh}, \tag{8}$$

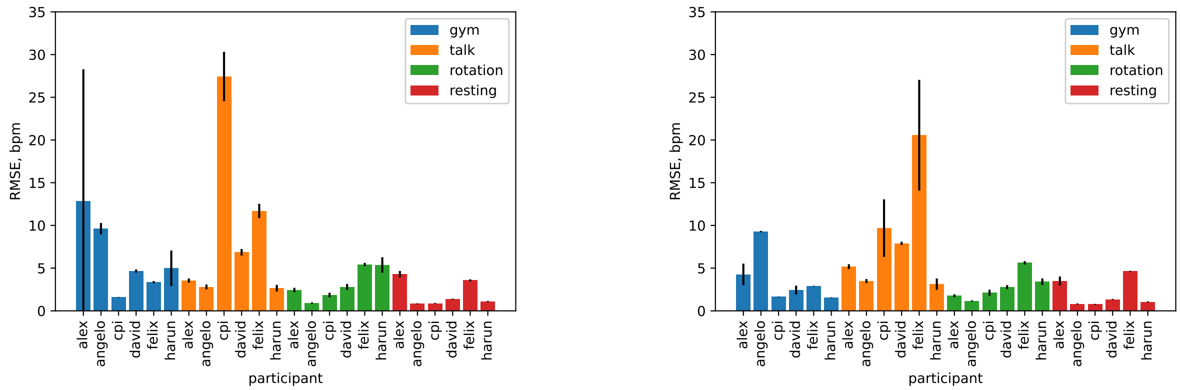
where N_{thresh} is threshold on effective number of particles. The final estimate is weighted sum of the samples

$$\hat{f}_t = \sum_{i=1}^m w_t^i f_t^i, \tag{9}$$

where m is the number of particles.

3.2. Initialization

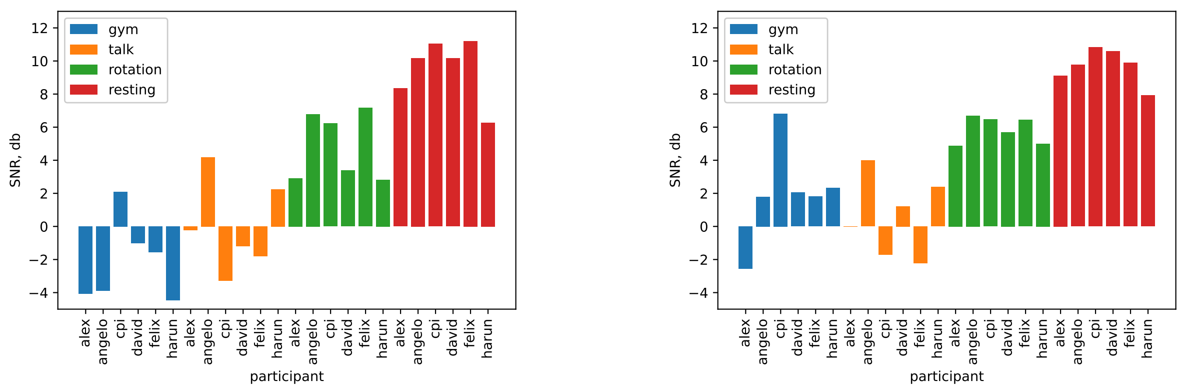
The straightforward way to initialize state of the particle filter would be the mode of power spectral density (PSD) obtained for the first window of BVP signal. This requires the person to be still and whole setup is not changing during this period of time, which does not hold true in some cases. For 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.



(a) LGI

(b) POS

Figure 2. Mean and standard deviation of RMSE for frequency tracker predictions on LGI dataset



(a) LGI

(b) POS

Figure 3. SNR of BVP signal for each video in LGI dataset

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-

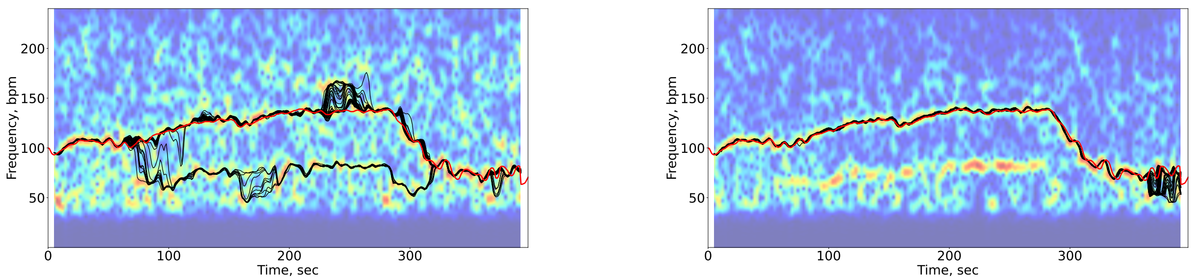


Figure 4. Periodogram of BVP signal for alex_gym video with heart rate predictions on it (black - 50 realizations of particle filter, red - ground truth)

cess noise was sampled from normal distribution $\mathcal{N}(0,\, \sigma_v^2)$, where σ_v was set to 0.5 bpm; standard deviation of measurement noise σ_u 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 that 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:

$$\frac{1}{AP}\sum_{i=1}^{A}\sum_{j=1}^{P} \underset{r}{max}(\underset{t}{M}(\hat{f}^{r}_{ijt} - f^{gt}_{ijt})), \tag{10}$$

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, $\hat{f}^{r}_{ijt}$ 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-

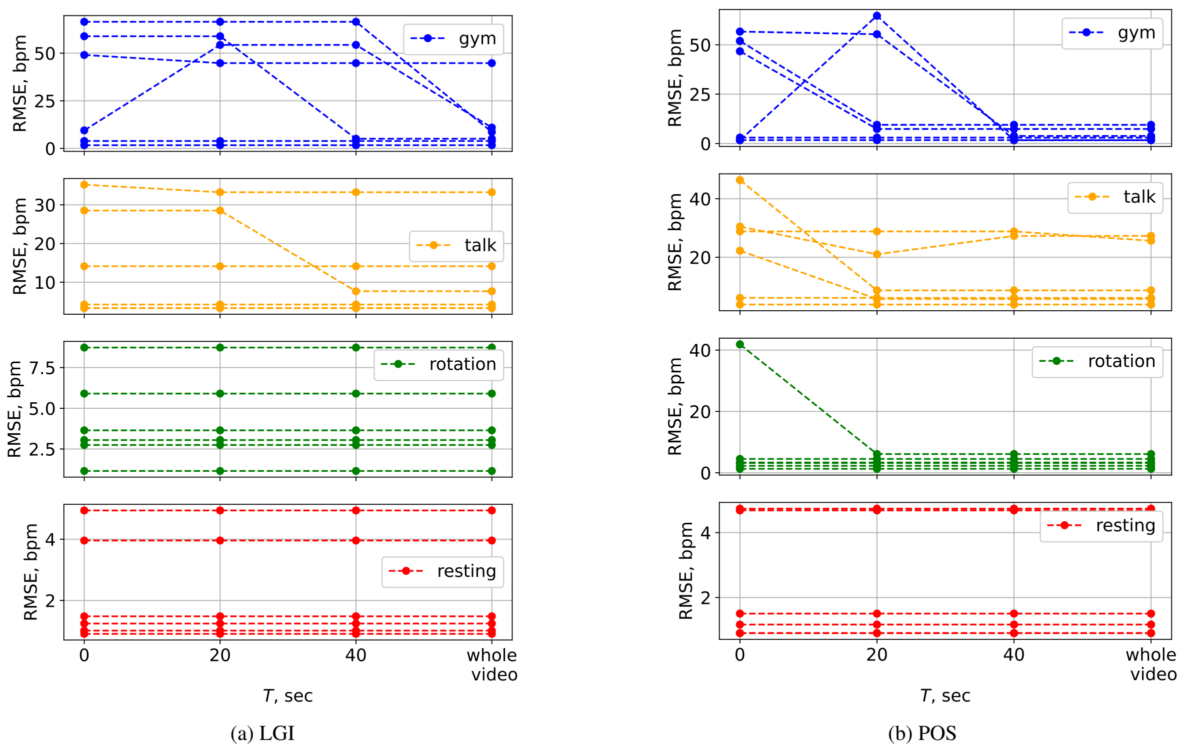


Figure 5. Evaluation results (Eq. (10)) depending on the duration of power accumulation step

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.

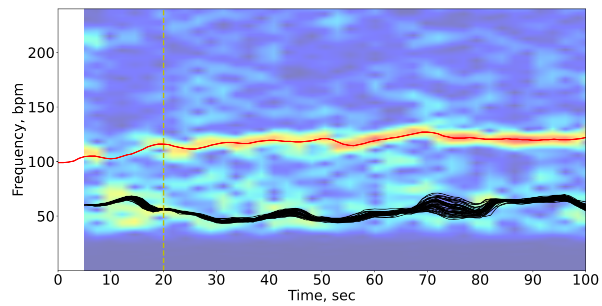
The algorithm performance depends on the type of noise, not only SNR. Though we obtain the worst SNR levels of BVP signal during gym session (Fig. 3), our algorithm is able to predict the right trace, since the heart rate track is preserved on the spectrogram and is separated from pedal rotation trace. 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 frequency 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.

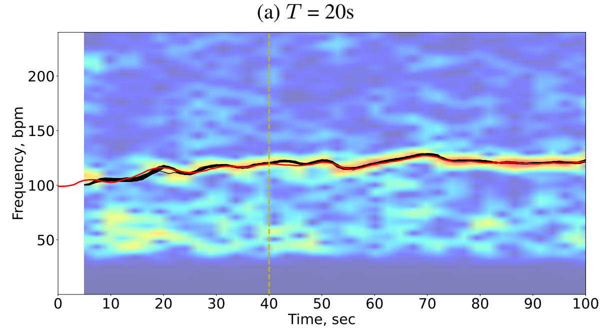
Finally, if $T = 0$ we do not use power accumulation scheme and select only one particle filter initialized with argmax frequency on the first 5 seconds of the video. Still, for POS restricting T to 40 seconds does not result in any noticeable deterioration comparing to the case when the trace is selected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



(a) $T = 20$s



(b) $T = 40$s

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 siganl 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

[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*, pages 1–1, 2020. 5

[2] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Edoardo Mortara. pyvhr: a python framework for remote photoplethysmography. *PeerJ Computer Science*, 8:e929, 2022. 5

[3] Constantino Alvarez Casado and Miguel Bordallo López. Face2ppg: An unsupervised pipeline for blood volume pulse extraction from faces. *arXiv preprint arXiv:2202.04101*, 2022. 1, 2, 5, 7

[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. 2

[5] Orchisama Das, JO Smith, and Chris Chafe. Real-time pitch tracking in audio signals with the extended complex kalman filter. In *Proceedings of the 20th International Conference on Digital Audio Effects*, pages 118–124, 2017. 3

[6] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2, 3

[7] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 1, 2

[8] Corentin Dubois, Manuel Davy, and Jérôme Idier. Tracking of time-frequency components using particle filtering. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 4, pages iv–9. IEEE, 2005. 3

[9] Jesse Fine, Kimberly L Branan, Andres J Rodriguez, Tananant Boonya-Ananta, Jessica C Ramella-Roman, Michael J McShane, and Gerard L Coté. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors*, 11(4):126, 2021. 1

[10] K Fujimoto, W Kasprzak, and N Hamada. Estimation and tracking of fundamental, 2nd and 3d harmonic frequencies for spectrogram normalization in speech recognition. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, pages 71–81, 2012. 3

[11] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3995–4004, 2021. 1, 2

[12] 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. 1, 2

[13] Fridolin Haugg, Mohamed Elgendi, and Carlo Menon. Effectiveness of remote ppg construction methods: A preliminary analysis. *Bioengineering*, 9(10):485, 2022. 1, 2

[14] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 383–389. IEEE, 2017. 2

[15] YungChien Hsu, Yen-Liang Lin, and Winston Hsu. Learning-based heart rate detection from remote photoplethysmography features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4433–4437. IEEE, 2014. 2

[16] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2018. 3

[17] Sunghan Kim, Anindya S Paul, Eric A Wan, and James McNames. Multiharmonic frequency tracking method using the sigma-point kalman smoother. *EURASIP Journal on Advances in Signal Processing*, 2010:1–13, 2010. 3

[18] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity.

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

[19] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 314–315, 2020. 6

[20] Xiaobai Li, Haomiao Sun, Zhaodong Sun, Hu Han, Antitza Dantcheva, Shiguang Shan, and Guoying Zhao. The 2nd challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2404–2413, 2021. 6

[21] Martin Lindfors. *Frequency tracking for speed estimation*. PhD thesis, Linköping University Electronic Press, 2018. 3

[22] 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. 1, 2

[23] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 5

[24] Yuriy Mironenko, Konstantin Kalinin, Mikhail Kopeliovich, and Mikhail Petrushan. Remote photoplethysmography: Rarely considered factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[25] Sharad Nagappa and James R Hopgood. Frequency tracking of biological waveforms. *Institute for digital communications, University of Edinburg, Edinburg*, 2006. 3, 4

[26] William Ng, Chunlin Ji, WK Ma, and Hing Cheung So. A study on particle filters for single-tone frequency tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 45(3):1111–1125, 2009. 3

[27] 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. 3

[28] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, pages 295–310. Springer, 2020. 3

[29] Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1254–1262, 2018. 5

[30] Žan Pirnar, Miha Finžgar, and Primož Podržaj. Performance evaluation of rppg approaches with and without the region-of-interest localization step. *Applied Sciences*, 11(8):3467, 2021. 1

[31] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2

[32] Ambareesh Revanur, Zhihua Li, Umur A Ciftci, Lijun Yin, and László A Jeni. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2760–2767, 2021. 6

[33] Frida Sandberg, Martin Stridh, and Leif Sornmo. Frequency tracking of atrial fibrillation using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 55(2):502–511, 2008. 3

[34] Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of deep learning based blood pressure prediction from ppg and rppg signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3820–3830, 2021. 1

[35] Vinothini Selvaraju, Nicolai Spicher, Ju Wang, Nagarajan Ganapathy, Joana M Warnecke, Steffen Leonhardt, Ramakrishnan Swaminathan, and Thomas M Deserno. Continuous monitoring of vital signs using cameras: A systematic review. *Sensors*, 22(11):4097, 2022. 1

[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, UK*, pages 3–6, 2018. 2

[37] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2

[38] Wenjin Wang, Albertus C den Brinker, and Gerard De Haan. Discriminative signatures for remote-ppg. *IEEE Transactions on Biomedical Engineering*, 67(5):1462–1473, 2019. 1, 2

[39] 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. 1, 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

[41] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1, 2

[42] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4186–4196, 2022. 1, 2

[43] Changchen Zhao, Peiyi 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-*

tional Conference on Computer Vision (ICCV) Workshops, Oct 2019. 1

[44] Qiang Zhu, Mingliang Chen, Chau-Wai Wong, and Min Wu. Adaptive multi-trace carving for robust frequency tracking in forensic applications. *IEEE Transactions on Information Forensics and Security*, 16:1174–1189, 2020. 3

[45] Qiang Zhu, Chau-Wai Wong, Chang-Hong Fu, and Min Wu. Fitness heart rate measurement using face videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2000–2004. IEEE, 2017. 2