NON-CONTACT PHOTOPLETHYSMOGRAM AND INSTANTANEOUS HEART RATE ESTIMATION FROM INFRARED FACE VIDEO

Natalia Martinez[⋆] *Martin Bertran*[⋆] *Guillermo Sapiro*[⋆] *Hau-Tieng Wu*[†]

[⋆] Department of Electrical and Computer Engineering, Duke University
[†] Department of Mathematics and Department of Statistical Science, Duke University

ABSTRACT

Extracting the instantaneous heart rate (iHR) from face videos has been well studied in recent years. It is well known that changes in skin color due to blood flow can be captured using conventional cameras. One of the main limitations of methods that rely on this principle is the need of an illumination source. Moreover, they have to be able to operate under different light conditions. One way to avoid these constraints is using infrared cameras, allowing the monitoring of iHR under low light conditions. In this work, we present a simple, principled signal extraction method that recovers the iHR from infrared face videos. We tested the procedure on 7 participants, for whom we recorded an electrocardiogram simultaneously with their infrared face video. We checked that the recovered signal matched the ground truth iHR, showing that infrared is a promising alternative to conventional video imaging for heart rate monitoring, especially in low light conditions. Code is available at `https://github.com/natalialmg/IR_iHR`.

1. INTRODUCTION

The gold standard for monitoring instantaneous heart rate (iHR) is electrocardiogram (ECG) [1]. Another popular non-invasive technique is photoplethysmogram (PPG) [2, 3]. Both techniques require direct skin contact with the subject, which might not be suitable in contexts such as driver drowsiness, or sleep monitoring. PPG relies on measuring the rapid variations in light absorption in an illuminated skin region caused by the difference in absorption curves for oxigenated and non-oxigenated blood. This principle motivated the use of digital cameras to measure the plethysmographic signals from face videos under ambient light conditions [4–6]. Several methodologies for estimating heart rate from face videos have been developed over the years [7–12]. In particular, [13] provides a comprehensive overview of the history of the research done in this area and compares the performance of some of these approaches. As a general rule, most of these methods need an illumination source, depend on color band manipulation, and require control over the signal acquisition process (e.g., controlled light sources, or subjects remaining motionless during acquisition).

The recent inclusion of infrared (IR) cameras in many conventional devices, coupled with their resilience to low-light and variable-light conditions, make them especially attractive for remote monitoring in the context of iHR detection. Their use has just now started to be explored in the detection of heart rate using infrared face videos [14, 15], but so far these approaches are limited to estimating a heart rate average over a considerable time frame (over 30 seconds). This paper shows that, under controlled motion conditions, it is feasible to extract even sub-second approximations to the iHR using basic spatiotemporal analysis and time-frequency analysis.

We describe this approach, and show its performance on face IR videos acquired using a Kinect camera from 7 healthy volunteers. The extracted iHR is compared against ECG and contact PPG ground truth signals that were simultaneously acquired.

2. NON-CONTACT PPG SIGNAL FROM IR VIDEO

Here we describe the proposed algorithm to construct the non-contact PPG signal from an IR face video and hence extract the instantaneous heart rate. We divide this process into three main steps. The first step is detecting and segmenting the face in the video into n_r disjoint spatial regions. Secondly, we take the mean activity of each region, denoise it, and decompose it into a smaller subset of sources. Finally, we introduce a signal quality index to select the signals of interest, and combine them to construct the non-contact PPG signal we are after. Figure 1 summarizes the initial preprocessing stages, while Figure 2 shows an example of the subsequent recovery process of the non-contact PPG.

2.1. Input IR video

For each subject, denote the recorded IR video as $V : \mathbb{R} \to \mathbb{R}^{n\times m}$, where $V(t)$ denotes the recorded frame at time t, which is of size $n \times m$ (height $\times$ width). Suppose the video is sampled every τ seconds; that is, sampled at $1/\tau$ Hz, and the recording starts at time 0 and lasts for T seconds. We have

WORK PARTIALLY SUPPORTED BY DOD, NIH, NSF, GOOGLE, CISCO, AND MICROSOFT.

thus $n_t = \lfloor T/\tau \rfloor$ frames. In this study, $1/\tau = 58Hz$. We additionally assume that the subject's head is fixed, so major movements between frames are ignored.

2.2. Preprocessing the IR video

We detect the boundaries of the face using the Dlib landmark detector [16] on the average face location, frame-by-frame detection is not performed since the subject is assumed to be immobile. An example is shown on Figure 1(a). We then divide the area inside the detected face into n_r disjoint regions following a predefined mesh grid.

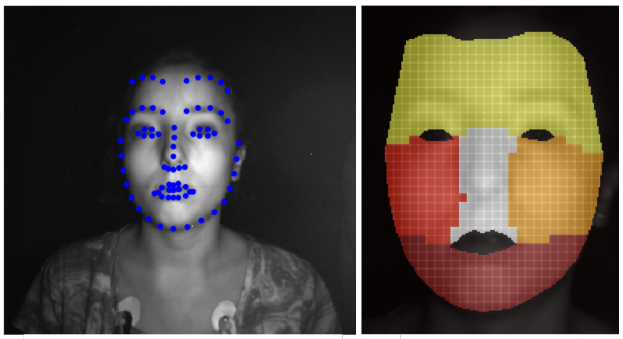
Denote those disjoint regions as R_i, $i = 1, \ldots, n_r$. For each video frame $V_j := V(j\tau) \in \mathbb{R}^{n \times m}$, $j = 1, \ldots, n_t$, we compute the mean IR value on each region R_i

$$y_{i,j} := \frac{1}{|R_i|} \sum_{(x,y) \in R_i} V_j(x, y).$$

As a result, we obtain the data matrix

$$Y_0 \in \mathbb{R}^{n_r \times n_t}.$$

In other words, the i-th row of matrix Y_0 contains a time series with the mean IR activity over region R_i, there are n_r such regions defined across the face. We will refer to these as channels. Note that the constructed data matrix is commonly encountered in spatiotemporal analysis. For the purposes of this study, the face is subdivided into regions using a non-overlapping 5×5-pixel grid. See Figure 1(b) for illustration.



(a) Detect facial landmarks on IR video. (b) Segment face into regions.

Fig. 1: Outline of video preprocessing pipeline. Figure a) shows how the facial landmarks are detected using Dlib. Figure b) shows how the detected face is subdivided using a 5x5 pixel grid; these regions can be grouped into 5 major facial areas. The mean activity signal of all grid elements compose our observation matrix Y_0.

To denoise the time series, we apply to each channel in the data matrix Y_0 an order 5 bandpass Butterworth with cut-off frequencies at 24 and 300 bpm, a range that comfortably acommodates most normal heart rates. Denote the filtered signals as the data matrix Y. This bandpass filter is chosen based on the physiological knowledge that, for a normal subject, the heart rate is between 40 and 200 bpm.

2.3. Low rank spatiotemporal model

We assume that the IR video captures different physiological dynamics, such as respiration, body movement, and hemodynamics, among others. Denote these physiological sources as $X_i \in \mathbb{R}^{n_t}$, where $i = 1, \ldots, n_s$ and $n_s \leq \min\{n_r, n_t\}$. Note that in general X_i and X_j might not be orthogonal when $i \neq j$; for example, the hemodynamics and respiration might be coupled due to the respiratory sinus arrhythmia.

The data matrix Y is then modeled as a mixture of these n_s source signals with additive and uncorrelated noise

$$Y = AX + \sigma Z, \tag{1}$$

where $X \in \mathbb{R}^{n_s \times n_t}$ contains the physiological source signals, $A \in \mathbb{R}^{n_r \times n_s}$ is the source mixture matrix, Z is a noise matrix with independent and identically distributed entries with zero mean, unit variance and finite fourth moment, and $\sigma^2 > 0$ is a scalar constant that describes the noise variance. In other words, the recorded signal on each region, Y_i, is a mixture of different sources via A, contaminated by noise. We further make the low rank assumption that n_s is fixed and small. This assumption means that there are limited sources of physiological dynamics that are captured by the IR video.

2.4. Determine important sources

Due to the low-rank assumption and the high-dimensional nature of the spatiotemporal model, apply SVD to the data matrix Y:

$$Y = U\Lambda V, \tag{2}$$

where $U \in O(n_r)$ consists of the left singular vectors, $V \in O(n_t)$ consists of the right singular vectors, and $\Lambda \in \mathbb{R}^{n_r \times n_t}$ consists of singular values $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_{\min\{n_r, n_t\}} \geq 0$. Denote u_i and v_i to be the i-th left and right singular vectors respectively. Note that V contains the relevant temporal signals that are mixed in each region, and U their weight in each spatial location. An example is illustrated in Figure 2.

Denote $\beta := n_t/n_s$ and denote n_s^* is the number of singular values such that $\eta(\sigma_i/\sigma) > 0$. Since the noise level σ is in general not known, we estimate it as proposed in [17]:

$$\hat{\sigma} = \frac{\text{median}(\vec{\sigma_i})}{\sqrt{\mu_b}},$$

where μ_b is the median of the Marcenko-Pastur distribution [18] with parameter $\lambda_\pm = (1 \pm \sqrt{\beta})^2$. Applying this procedure to Y reduced the number of non-zero singular values by over 80% on average.

2.5. Reconstructing the non-contact PPG signal

Due to the non-orthogonal nature of physiological sources, we cannot recover X directly from Y by applying the usual blind source separation technique. We thus propose the following procedure to reconstruct the non-contact PPG signal.

Define a signal quality index (SQI) for a signal x of length n_t as

$$Q(x) := \frac{\int_{\frac{3}{4}f_p}^{\frac{5}{4}f_p} |\hat{x}(f)|df}{\int_{\frac{1}{2}f_p}^{2f_p} |\hat{x}(f)|df},$$

where $\hat{x}$ is the Fourier transform of the time series x, and f_p is the expected heart rate of a normal subject. Note that $Q(x)$ quantifies how concentrated the time series x is around f_p in the frequency domain.

We rank all temporal signals v_i, where $i = 1, \ldots, n_s^*$, according to their SQIs $Q(v_i)$. Consider the reordering permutation $q : \{1, \ldots, n_s^*\} \to \{1, \ldots, n_s^*\}$ so that $Q_{q(1)} \geq Q_{q(2)} \geq \ldots$. Our hemodynamic estimator, the non-contact PPG signal denoted as PPG_{IR}, is defined as

$$\text{PPG}_{\text{IR}} := \sum_{i=q(1)}^{q(J)} v_i \in \mathbb{R}^{n_t},$$

for $J \in \mathbb{N}$ chosen by the user. Here we determine J by greedily accumulating the sources until the maximal quality is achieved; that is,

$$q(J) = \arg\max_j Q\left(\sum_{i=q(1)}^{q(j)} v_i\right).$$

Figure 2 shows an outline of the recovery process for non-contact PPG over the full face. Figure 3 shows the recovered iHR signal when we applied the proposed method to the channels contained in each of the five major facial areas independently, this is provided merely for illustration purposes. In general, using the entire facial area provided the best results. Figure 4 shows short time segments of non-contact PPG compared against ground truth contact PPG. Additional examples will be provided in the following sections.

2.6. Estimation of the instantaneous heart rate

Denote the short time Fourier transform (STFT) of the constructed non-contact PPG signal as $S_{\text{PPG}_{\text{IR}}} \in \mathbb{C}^{n_t \times (n_t/2)}$, where $S_{\text{PPG}_{\text{IR}}}(t, f)$ is the STFT coefficient at time t/τ and frequency f/T. From the STFT we extract the dominant curve using the curve extractor proposed in [19],

$$\hat{c} = \arg\max_{c \in \mathbb{N}^{n_t}} \sum_{t=1}^{n_t} \log |S_{\text{PPG}_{\text{IR}}}(t, c(t))| \qquad (3)$$
$$- \lambda \sum_{t=2}^{n_t} |c(t) - c(t-1)|) \in \mathbb{N}^{n_t},$$

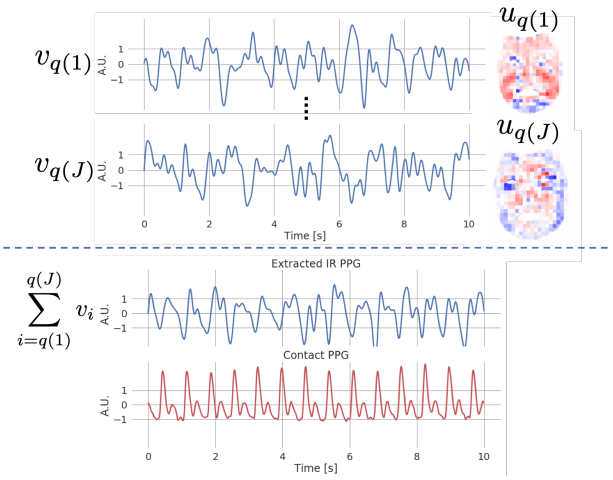


Fig. 2: Top half of figure shows left (u) and right (v) singular vectors sorted by SQI in descending order. The resulting accumulated non-contact PPG (PPG_{IR}) is shown on the bottom, ground truth contact PPG is shown for comparison. Contact and noncontact PPG show well matched cycles.

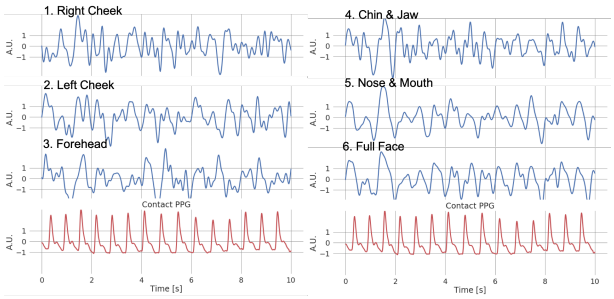


Fig. 3: Results of applying the proposed framework to the channels contained in each of the 5 major facial areas independently. All areas contain a measure of iHR information, the best results are obtained by analyzing the full face as a whole. Ground truth PPG is provided for comparison.

where $\lambda > 0$ is a regularization constant. The iHR is thus determined by

$$\text{iHR} := \hat{c}/T \in \mathbb{R}^{n_t}.$$

Figure 5 shows the obtained PPG_{IR} and its STFT; ground truth iHR from ECG is also shown for comparison.

3. EXPERIMENTS

We acquired 9 simultaneous ECG, PPG, and IR face video using a standard patient monitor (Philips IntelliVue MP70 Patient Monitor) and a Microsoft Kinect camera. The clocks in the Kinect camera and the patient monitor were synchronised with a time accuracy of $\pm 1s$. Acquisitions were done over 7 healthy subjects. The subjects were asked to look straight into the camera and maintain a steady posture, but otherwise

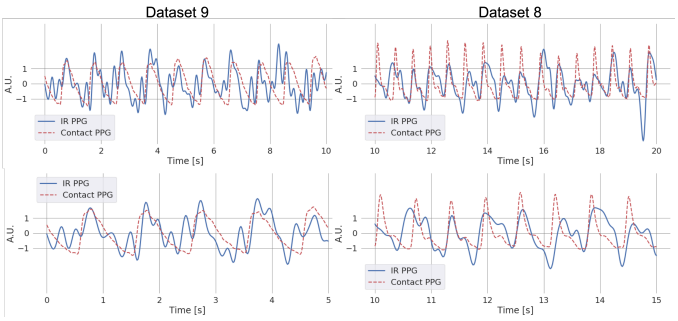


Fig. 4: Estimated PPG_{IR} (solid blue) compared against ground truth contact PPG (dashed red) at $10s$ and $5s$ timescales for two datasets. Overall, cycles are well matched between contact and noncontact PPG. Datasets are taken from subjects with dissimilar resting heartrate.

behave, blink, and breathe normally. The instantaneous heart rate (iHR) was estimated from the IR video using the process described in Section 2. Ground truth iHR was extracted from the ECG signal using the R-peak detection algorithm implemented in the python library biosppy [20].

4. RESULTS

For each of the 9 datasets we measured the differences between the recovered iHR signal and ground truth using root mean square error (RMSE) and relative error

$$\frac{1}{n_t}\sum_{t=1}^{n_t}\frac{|iHR(t)-iHR_{ECG}(t)|}{iHR_{ECG}(t)}.$$

Table 1 shows these values. Figure 5 shows the extracted iHR signals. Implemented code is available at `https://github.com/natalialmg/IR_iHR`.

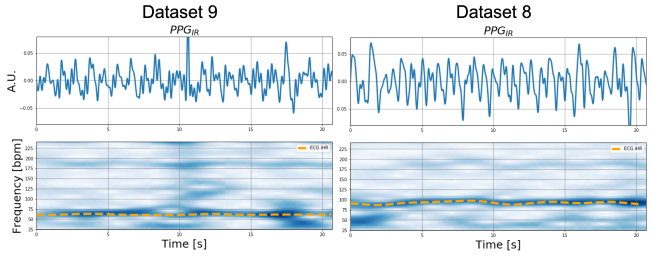


Fig. 5: Recovered noncontact PPG (PPG_{IR}) and its corresponding STFT are shown for dataset 9 and 8. Dotted orange line shows the ground truth iHR recovered from ECG signal. Spectrograms show good correspondence between recovered IR iHR and ground truth iHR.

In general, RMSE results averaged for longer time-frames ($30s$) are satisfactory. Perhaps surprisingly, RMSE results for iHR at $1s$ intervals are also reasonable. Figure 5 shows good correspondence between the ground truth ECG iHR and the STFT of the recovered non-contact PPG.

Dataset / Subject	RMSE [bpm]			Relative error [%]
	Every 1s	Every 10s	Every 30s	Every 30 s
d1/1	5.39	4.27	4.03	4.50
d2/1	5.61	4.99	5.22	6.51
d3/2	4.71	4.44	3.70	4.40
d4/2	3.59	2.87	1.33	1.56
d5/3	4.39	3.86	1.95	2.60
d6/4	4.95	4.65	2.91	3.58
d7/5	2.21	1.31	1.02	1.60
d8/6	3.30	1.42	0.23	0.25
d9/7	2.38	1.26	0.66	1.08

Table 1: Error measures across datasets

5. CONCLUDING REMARKS

In this paper, we extracted non-contact PPG from IR facial video. We showed that a simple, principled method based on matrix decomposition was sufficient to recover instantaneous heart rate with small relative errors on a second-by-second basis when subjects remain relatively stationary.

This suggests the viability of IR for non-contact PPG, particularly when we consider the low-light and varying-light performance of IR in general compared to traditional RGB methods. Additional work is required to adequately and robustly correct for motion artifacts. Improvements can be done on the process by which we combine the singular vectors to obtain our final hemodynamic estimator. Finally, more research needs to be done on the characterization of absorption curves of biological processes of interest in the near infrared spectrum. We leave this physiological research as a future collaborative work. We could also consider more sophisticated time-frequency representation tools to further analyze the obtained non-contact PPG signal for the instantaneous heart rate estimation. A more general manifold learning algorithm and matrix denoise technique can be applied to capture motion and time latency; for example, due to the high dimensional nature of Y, the matrix Y can be denoised by the optimal shrinkage algorithm proposed in [21]: $\tilde{Y} = \sum_{i=1}^{n_s^*}\sigma\eta(\sigma_i/\sigma)u_i v_i^T$, where $\eta(y) = \begin{cases}\frac{\sqrt{(y^2-\beta-1)^2-4\beta}}{y} & y > 1+\sqrt{\beta} \\ 0 & y \le 1+\sqrt{\beta}\end{cases}$ is the optimal shrinkage under the Frobenius norm. This approach has the potential to further improve the overall quality of the signal. We will explore these possibilities in future work.

6. REFERENCES

[1] JA Dawson, COF Kamlin, C Wong, AB Te Pas, M Vento, TJ Cole, SM Donath, SB Hooper, PG Davis, and CJ Morley, "Changes in heart rate in the first minutes after birth," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 95, no. 3, pp. F177–F181, 2010.

[2] Aymen A Alian and Kirk H Shelley, "Photoplethysmography," *Best Practice & Research Clinical Anaesthesiology*, vol. 28, no. 4, pp. 395–406, 2014.

[3] John Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, pp. R1, 2007.

[4] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[5] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[6] Maria I Davila, Gregory F Lewis, and Stephen W Porges, "The physiocam: cardiac pulse, continuously monitored by a color video camera," *Journal of Medical Devices*, vol. 10, no. 2, pp. 020951, 2016.

[7] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[8] Sungjun Kwon, Hyunseok Kim, and Kwang Suk Park, "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 2174–2177.

[9] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.

[10] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal, "Distanceppg: Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[11] Antony Lam and Yoshinori Kuno, "Robust heart rate measurement from video using select random patches," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3640–3648.

[12] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2396–2404.

[13] Chen Wang, Thierry Pun, and Guillaume Chanel, "A comparative survey of methods for remote heart rate detection from frontal face videos," *Frontiers in Bioengineering and Biotechnology*, vol. 6, 2018.

[14] Jie Chen, Zhuoqing Chang, Qiang Qiu, Xiaobai Li, Guillermo Sapiro, Alex Bronstein, and Matti Pietikäinen, "Realsense= real heart rate: Illumination invariant heart rate estimation from videos," in *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*. IEEE, 2016, pp. 1–6.

[15] Qi Zhang, Yimin Zhou, Shuang Song, Guoyuan Liang, and Haiyang Ni, "Heart rate extraction based on near-infrared camera: Towards driver state monitoring," *IEEE Access*, vol. 6, pp. 33076–33087, 2018.

[16] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[17] David L Donoho and Matan Gavish, "The optimal hard threshold for singular values is 4/ 3," *arXiv preprint*, 2013.

[18] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur, "Distribution of eigenvalues for some sets of random matrices," *Matematicheskii Sbornik*, vol. 114, no. 4, pp. 507–536, 1967.

[19] Antonio Cicone and Hau-Tieng Wu, "How nonlinear-type time-frequency analysis can help in sensing instantaneous heart rate and instantaneous respiratory rate from photoplethysmography in a reliable way," *Frontiers in Physiology*, vol. 8, pp. 701, 2017.

[20] Carlos Carreiras, Ana Priscila Alves, André Lourenço, Filipe Canento, Hugo Silva, Ana Fred, et al., "BioSPPy: Biosignal processing in Python," 2015–, [Online; accessed Jan 29, 2019].

[21] Matan Gavish and David L Donoho, "Optimal shrinkage of singular values," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2137–2152, 2017.