The Benefit of Distraction: Denoising Camera-Based Physiological Measurements using Inverse Attention: Supplementary Materials

Anonymous ICCV submission

Paper ID 3921

1. Details About the Datasets

AFRL [3]: 300 videos of 25 participants were recorded as 658×492 pixel images at 120 fps. Fingertip reflectance photoplethysmograms (PPG), electrocardiograms (ECG), and breathing signals were recorded as ground truth signals. We used the ECG signals to compute the HR estimation errors, the PPG signals to train the network for estimating HR, and breathing signals for computing the errors and training the network for BR estimation. Each participant was recorded 12 times in each five-minute experiment with varying motion and two different backgrounds. The participants: 1) sat still and rested their chin on a headrest, 2) sat still without the headrests, 3) moved their head horizontally at a speed of 10 degrees/second, 4) 20 degrees/second, 5) 30 degrees/second, 6) reoriented their head randomly once every second. We center-cropped the ARFL video frames to 492×492 pixels to remove the blank background areas.

MMSE-HR [10]: 102 videos of 40 participants were recorded at 25 fps capturing 1040×1392 resolution images during spontaneous emotion elicitation experiments. Ground truth blood pressure (BP) wave was measured at 1000 fps and average HR updated after every heart beat. We used the blood pressure waves to train the network and the average HR to compute the HR estimation errors. 19 videos had erroneous average HR estimates, so we recomputed them by using the BP waveform. We detected peaks in the blood pressure waveform and computed the inter-beat interval (IBI) between the detected peaks. Heart rate is estimated as $\frac{1}{\mu(IBI)}$ where $\mu(IBI)$ is the mean IBI. This dataset is more challenging than the AFRL dataset [3] because of the sudden facial motions and rapidly changing heart rate during the experiments.

MR-NIRP (NIR) [5]: Eight participants were recorded with a NIR camera. The videos were recorded at 640×640 resolution and 30 fps. Fingertip transmission photoplethysmograms were recorded as ground truth signals. Each participant was recorded twice, once sitting still and once performing motion tasks involving talking and randomly moving the head. Because the background in MR-NIRP was not uniform, we applied face detection in the first video frame and cropped a rectangular region with 110% width and height of the detected bounding box. This dataset is particularly challenging because the physiological signals are very weak in NIR and are difficult to recover in presence of head motion [4, 8].

2. Error Metrics

To evaluate the performance of our proposed approach we used the following four standard error measures (MAE, RMSE, Correlation, SNR), and we defined a new measure (Waveform MAE) to measure the waveform dynamics.

Mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{i=1}^{N} |R_i - \widehat{R}_i|}{N} \tag{1}$$

where N is the total number of time windows, R_i is the ground truth heart rate (HR) measured with a contact sensor for each 30 second time window and $\widehat{R}_i$ is the estimated HR from the video.

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (R_i - \widehat{R}_i)^2}{N}} \tag{2}$$

Pearson's Correlation Coefficient (ρ): computed between HR estimates from each time window $\widehat{R} = [\widehat{R})(1), ..., \widehat{R}(N)]$ and the ground truth HR measurements $R = [R(1), ..., R(N)]$.

Signal-to-noise ratio (SNR): calculated as the ratio of the area under the curve of the power spectrum around the first and the second harmonic of the ground truth HR frequency divided by the rest of the power spectrum within the physiological range of 42 to 240 bpm [2]:

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{42}^{240}((U_t(f)S(f))^2}{\sum_{42}^{240}((1-U_t(f))S(f))^2}\right) \quad (3)$$

where S is the power spectrum of the estimated iPPG signal, f is the frequency in beats per minute (BPM) and $U_t(f)$ is equal to one for frequencies around the first and second harmonic of the ground truth HR (HR-6 bpm to HR+6 bpm and 2*HR-6 bpm to 2*HR+6 bpm), and 0 everywhere else.

Waveform Mean Absolute Error (WMAE):

$$\text{WMAE} = \frac{\sum_{i=1}^{N}|W_i - \widehat{W}_i|}{N} \quad (4)$$

where W_i is the ground truth pulse waveform obtained with the contact sensor for each 30 second time window and $\widehat{W}_i$ is the estimated pulse waveform from the video.

3. Implementation of Baseline Methods

We compared the performance of our proposed approach to state-of-the-art supervised method using a convolutional attention network (CAN) and several unsupervised methods. We implemented the CAN, CHROM, ICA, and POS methods to evaluate them on the datasets we used.

For the CHROM, ICA and POS methods face detection was first performed using MATLAB's face detection (`vision.CascadeObjectDetector()`). This was fixed for all methods, to avoid the influence of the face detector on performance.

CHROM [2]. This method uses a linear combination of the chrominance signals obtained from the RGB video. The $[x_R, x_G, x_B]$ signals are filtered using a zero-phase, 3rd-order Butterworth bandpass filter with pass-band frequencies of [0.7 2.5] Hz. Following this, a moving window method of length 1.6 seconds (with overlapping windows and a step size of 0.8 seconds) is applied. Within each window the color signals are normalized by dividing by their mean value to give $[\bar{x_r}, \bar{x_g}, \bar{x_b}]$. These signals are bandpass filtered using zero-phase forward and reverse 3rd-order Butterworth filters with pass-band frequencies of [0.7 2.5] Hz. The filtered signals $[\mathrm{y}_r, \mathrm{y}_g, \mathrm{y}_b]$ are then used to calculate S_{win}:

$$S_{win} = 3(1-\frac{\alpha}{2})y_r - 2(1+\frac{\alpha}{2})y_g + \frac{3\alpha}{2}y_b \quad (5)$$

Where α is the ratio of the standard deviations of the filtered versions of A and B:

$$A = 3y_r - 2y_g \quad (6)$$

$$B = 1.5y_r + y_g - 1.5y_b \quad (7)$$

The resulting outputs are scaled using a Hanning Window and summed with the subsequent window (with 50% overlap) to construct the final blood volume pulse (BVP) signal.

ICA [6]. This approach involves spatial averaging the pixels by color channel in the region of interest (ROI) for each frame to form time varying signals $[\mathrm{x}_R, \mathrm{x}_G, \mathrm{x}_B]$. Following this, the observation signals are detrended. A Z-transform is applied to each of the detrended signals. The Independent Component Analysis (ICA) (JADE implementation) is applied to the normalized color signals.

POS [9]. The intensity signals $[x_R, x_G, x_B]$ are computed for each camera channel. A moving window of length 1.6 seconds (with overlapping windows and with a step size of one frame), is applied. For each time window, the signal is divided by its mean to give $[\bar{x_r}, \bar{x_g}, \bar{x_b}]$. Following this, X_s and Y_s are calculated where:

$$X_s = \bar{x}_g - \bar{x}_b \quad (8)$$

$$Y_s = -2\bar{x}_r + \bar{x}_g + \bar{x}_b \quad (9)$$

X_s and Y_s are then used to calculate S_{win}, where:

$$S_{win} = X_s + \frac{\sigma(X_s)}{\sigma(Y_s)}Y_s \quad (10)$$

The resulting outputs of the window-based analysis are used to construct the final BVP signal in an overlap-add fashion.

CAN [1]. The supervised convolutional attention neural network is described in detail in the main text [1]. Following the implementation in that paper, we did not use face detection but rather we pass the full frame to the network after cropping the center portion to make the frame a square with W=H.

Signal Pre-processing. We bandpass filtered the physiological signals and corruption estimates to 0.7 Hz - 2.5 Hz range and detrended them [7] before feeding them into the LSTM. We set the detrending parameter λ for each dataset based on the video frame rate (λ = 500 for AFRL [3] and λ = 50 for MMSE-HR [10] and MR-NIRP [5]). We normalized the signals and corruption estimates with AC/DC normalization by subtracting the temporal mean and dividing by the temporal standard deviation computed for each video. We additionally normalized the amplitude range of the signals, corruption estimates, and the ground truth signals to -1 and 1. Finally, we resampled all sequences to 30 fps.

4. Statistical Significance

We computed F-tests to verify that our errors had significantly lower variance (spread) than the baselines. For AFRL and MR-NIRP which had longer videos, we computed the error metrics for each video, and for the shorter

MMSE-HR, we computed them for all time windows in the dataset. In addition to lower mean errors, for all datasets, our approach led to a significantly lower spread in the HR and BR MAE and RMSE. AFRL (300 videos): HR MAE: F = 0.54, $p < 0.01$, HR RMSE: F = 0.56, $p < 0.01$, BR MAE: F = 0.36, $p < 0.01$, BR RMSE: F = 0.48, $p < 0.01$, MMSE-HR (131 windows): MAE: F = 0.26, $p < 0.01$, RMSE: F = 3.92 $p < 0.01$, MR-NIRP (15 videos): MAE F = 7.94 $p < 0.01$, RMSE F = 6.63, $p < 0.01$.

5. Comparison of Corruption Estimation

Corruption Definition. We compared the performance of our proposed denoising framework with corruption channels computed from a single red, green or blue camera channel to using all three R, G, B channels. We hypothesized that the blue channel might be the best one for the corruption representation for the physiological signals because the hemoglobin present in blood has the lowest absorption in the blue light spectrum and its intensity variations would be the least related to blood flow. Conversely, the green channel could also be a useful corruption representation, because it would contain information most similar to the physiological signals since the hemoglobin has the largest absorption in the green spectrum. However, we found that there is not a large difference between using any one of the single channels or all three channels. We report the detailed results in Table 1 on the AFRL dataset [3].

Inverse Mask Definition. We also compared computing the corruptions using a binary and a continuous inverse attention mask. The continuous mask was computed as a matrix of continuous values in which each element of the inverse mask M, $M_{i,j}$, was 1 - $A_{i,j}$ where A is the attention mask weights normalized from 0 to 1. The binary mask was computed by thresholding these values, where $A'_{i,j}$ = 1, if $A_{i,j}$ >T, where T is a threshold from 0 to 1. We found that we obtained comparable results with the binary and continuous masks as shown in Table 1.

Table 1. Different Inverse Mask Definitions on AFRL [3]. There was no systematic benefit of using R, G, B or RGB inputs or using the binary vs. continuous mask. We used the binary mask with RGB inputs for the results shown in the main paper.

	AFRL (All Tasks) [3]				
Method	MAE	RMSE	SNR	ρ	WMAE
Ours (LSTM RGB Binary Mask)	2.25	5.68	6.44	0.87	0.21
Ours (LSTM Red Binary Mask)	2.09	5.19	6.70	**0.89**	0.21
Ours (LSTM Green Binary Mask)	**2.04**	**5.11**	6.84	**0.89**	0.21
Ours (LSTM Blue Binary Mask)	2.18	5.27	6.59	0.88	0.21
Ours (LSTM RGB Continuous Mask)	2.10	5.61	**7.11**	0.87	**0.20**

Different Distraction Regions. We compared separately using corruption estimates from distraction regions closer to the face ("Center" of the frames) and further from the face ("Edges" of the frames). We used an LSTM model trained on all ignored regions for this experiment. When the motion was small, all regions contributed similarly to denoising. But when there was large head motion, regions close to the head (center of the frames) helped the most. See Table 2.

Figure 1. Comparison of attention masks and inverse attention masks on a video with and without glasses.

Table 2. Different Distraction Regions on AFRL [3]

	MAE						BVP SNR					
Region	1	2	3	4	5	6	1	2	3	4	5	6
Edges	**1.07**	**2.10**	1.92	2.10	2.68	8.74	**10.52**	7.23	8.59	6.04	3.07	-5.83
Center	1.08	2.11	**1.75**	**2.00**	**2.43**	**6.53**	10.50	**7.28**	**8.72**	**6.33**	**3.89**	**-4.47**

Performance on Subjects with Glasses. We compared the performance of our denoising approach and the baseline CAN method on subjects with and without glasses. We found that our method offers the largest improvements on subjects with glasses, as shown in Table 3. However, the attention masks output by CAN on subjects with and without glasses were comparable, as shown in Figure 1. Nine of the 25 subjects in the AFRL dataset were wearing glasses. No subjects in the MMSE-HR or MR-NIRP datasets were wearing glasses.

Table 3. Subjects with Glasses from AFRL [3]

Method	MAE	RMSE	SNR	ρ	WMAE
Ours (LSTM) with Glasses	**2.17**	**4.55**	**7.33**	**0.87**	0.21
CAN with Glasses	3.33	6.56	3.80	0.76	0.24
Ours (LSTM) no Glasses	2.55	5.79	4.68	0.59	**0.20**
CAN no Glasses	2.57	5.13	2.50	0.66	0.22

References

[1] 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

[2] 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] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 1, 2, 3

[4] Luis F Corral Martinez, Gonzalo Paez, and Marija Strojnik. Optimal wavelength selection for noncontact reflection photoplethysmography. In *22nd Congress of the International Commission for Optics: Light for the Development of the World*, volume 8011, page 801191. International Society for Optics and Photonics, 2011. 1

[5] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: towards driver mon-

itoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309. IEEE, 2018. 1, 2

[6] Ming-Zher Poh, Daniel McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2

[7] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002. 2

[8] Vytautas Vizbara. Comparison of green, blue and infrared light in wrist and forehead photoplethysmography. *BIOMEDICAL ENGINEERING 2016*, 17(1), 2013. 1

[9] 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

[10] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 1, 2