Beat-to-Beat Cardiac Pulse Rate Measurement From Video

Brian L. Hill
University of California, Los Angeles
Los Angeles, CA, USA
brian.l.hill@cs.ucla.edu

Xin Liu
University of Washington
Seattle, WA, USA
xliu0@cs.washington.edu

Daniel McDuff
Microsoft
Redmond, WA, USA
damcduff@microsoft.com

Abstract

Non-invasive cardiac sensing has many applications. Cameras specifically are ubiquitous, low-cost, spatial sensors that can also be used to capture context alongside physiological signals. However, sufficient precision is necessary for this technology to have an impact and for it to be trusted. Benchmark datasets and competitions have contributed significantly to advancing the state-of-the-art methods and improving transparency. We present an entry to the vision for vitals (V4V) challenge.

1. Introduction

Interest in the use of cameras to measure cardiopulmonary signals, including the cardiac pulse, has grown tremendously in recent years. Camera-based measurement has several attractive properties, as cameras are ubiquitous, low-cost, and can perform spatial and concomitant measurements without touching the body. Perhaps just as significantly, videos can be used to capture context that is difficult to obtain from a wearable device, including the identity of a subject, their non-verbal cues (e.g., facial expressions), position/posture, what activity they are performing, their environment, and their appearance (e.g., build, height, etc.).

Foundational work has demonstrated that digital images can be used to measure cardiac information from subtle pixel changes in videos of the human body [22, 1, 17, 20]. Computational methods aimed at making these measurements more robust were subsequently developed, many initially using unsupervised approaches [14, 4, 21, 18]. However, the performance of these methods has been superseded by supervised training [3, 23, 8, 9]. Neural models are typically able to learn more complex spatial and temporal relationships from video data. However, a large and diverse training set is necessary if these models are to generalize well.

Collecting training videos with the necessary diversity to develop models that generalize to new datasets is not trivial. The existing public imaging PPG datasets ([5, 24, 12, 2, 11, 13]) were each collected in a single environment. Therefore, all videos have similar background, lighting conditions, and position of the subjects relative to the camera. Synthetic data is one way to address this. High-fidelity simulations can be used to create videos with varied appearance, backgrounds, lighting conditions, motion, and facial expressions [10].

In the field of computer vision, benchmarks and challenges have formed a significant contribution to the research community [16, 19]. These efforts help establish a clear picture of the performance of different algorithms and also reveal limitations that can inspire future research. The remainder of this paper will summarize: 1) the model used for video-based pulse rate measurement, 2) the training procedure and datasets used, and 3) the results on the V4V challenge data [15].

2. Hybrid-CAN-RNN

The Hybrid-CAN architecture [9] has been shown to be highly effective at estimating pulse rate from video data, and achieved state-of-the-art results across a multitude of tasks. In the Hybrid-CAN architecture, the model is divided into two branches: (1) an appearance branch, which learns features from an RGB frame averaged over all time points in a window, and (2) a motion branch, which extracts features from the differences between consecutive video frames. Spatial attention mechanisms are used to share information between the two branches, encouraging the model to focus on regions of the image that contain useful signal (i.e. participant's skin) and ignore noisy regions (i.e. background).

The Hybrid-CAN architecture leverages local changes in both space and time. However, due to the convolutional nature of the model, learning a longer-term time-based representation is difficult to achieve. Therefore, to better model the temporal aspect of the PPG waveform, we extend the Hybrid-CAN architecture by integrating recurrent neural network (RNN) layers on top of the convolutional layers (Hybrid-CAN-RNN). Specifically, we apply 3D average pooling to the final layer of learned CNN features and in-

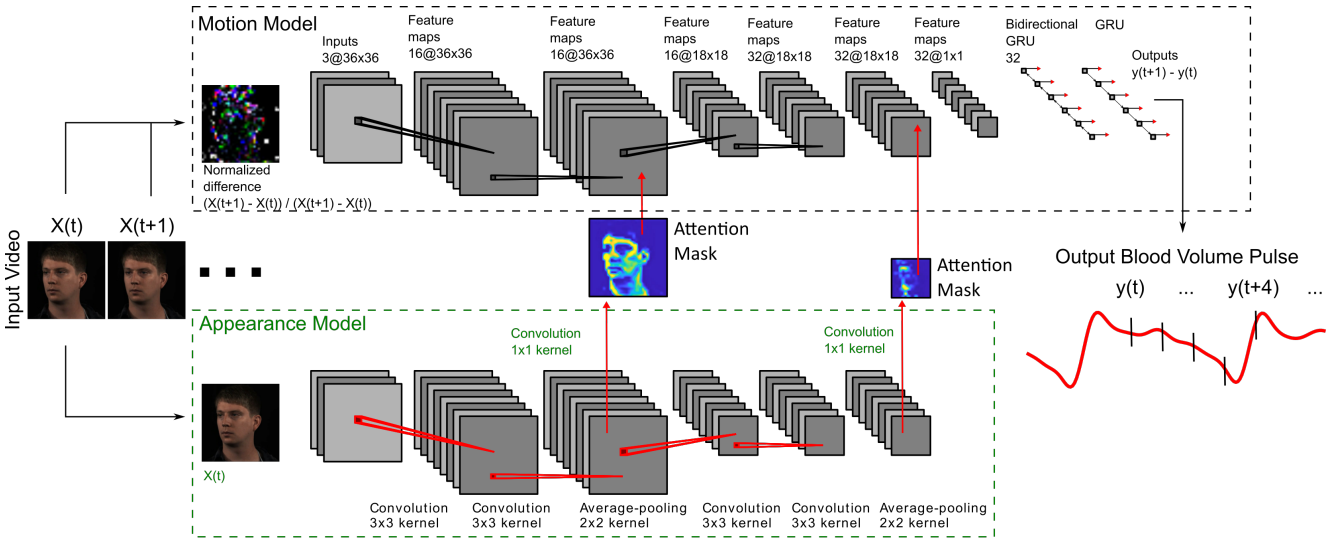

Figure 1. We present a novel neural architecture called Hybrid-CAN-RNN that leverages 3D Convolutions to perform spatial-temporal modeling and bi-directional GRU layers to predict high-quality time-series pulse waveform.

corporate a bidirectional gated recurrent unit (GRU) layer followed by an additional GRU layer to output the predicted the waveform signal at each time step. Figure 1 depicts the Hybrid-CAN-RNN architecture. This approach allows the 3D convolutional layers to extract features that are local in both the space and time dimensions, and the RNN layers to model the longer temporal transitions based on the representation learned by the 3D convolutional layers.

3. Datasets

AFRL [5] The AFRL dataset includes a total of 300 videos from 25 participants (17 male and 8 female). Each video in the dataset has a resolution of 658x492, and the sampling rate is 30 Hz. Gold-standard PPG signals were recorded using a contact reflective PPG sensor attached to the subject's index finger. Each participant was instructed to perform three head motion tasks including rotating the head along the horizontal axis, rotating the head along the vertical axis, and rotating the head randomly once every second to one of nine predefined locations. For horizontal and vertical head motion tasks, the subjects were instructed to conduct head motions with an increasing speed (10 degrees/second, 20 degrees/second, 30 degrees/second, 40 degrees/second, 50 degrees/second, 60 degrees/second).

BP4D+ [24] The BP4D dataset has 140 videos from 140 participants (82 female and 58 male) with ages ranging from 18 to 66 years old. The dataset also contains diverse racial ancestries including Black, White, Asian (both East-Asian and Middle-East-Asian), Hispanic/Latino, and others (e.g., Native American). The sampling rate used in the videos is 25 Hz, and the raw resolution is 1040 × 1392. The ground-truth blood pressure waveform was collected by a Biopac MP150 data acquisition system.

UBFC [2] The UBFC dataset has 42 videos from 42 participants. Each video has a resolution of 640x480 and the sampling rate is 30 Hz in uncompressed 8bit RGB format. The reference PPG signal was collected using a CMS50E transmissive pulse oximeter. The experiments were conducted in different indoor illumination and sunlight conditions.

Synthetics [10] To improve model generalization, we leverage recent work that uses highly-parameterized synthetic avatars to generate videos containing a diverse set of simulated subjects, movements, and backgrounds [10]. Using physiological waveforms signals from the MIMIC Physionet [6] database, we randomly sampled windows of PPG and respiration from real patients. The physiological waveform data were sampled to maximize examples from different patients. These waveforms were then used to drive the synthetic avatars' appearance. Specifically, the PPG signal is used to manipulate the base skin color and the subsurface radius [10]. The subsurface scattering is spatially weighted using an artist-created subsurface scattering radius texture which captures variations in the thickness of the skin across the face. Using the synthetic avatar pipeline, we generated 2800 6-second videos, where half of the videos were generated using hand-crafted facial motion/action signals, and the other half using facial motion/action signals extracted using landmark detection on real videos.

4. Training Details

We trained our model using a large dataset consisting of participants from the AFRL, UBFC, BP4D+ (including MMSE and V4V training datasets), in addition to the gen-

Table 1. Beat-to-Beat Pulse Rate Prediction on V4V Dataset

Training Data	MAE	RMSE	ρ
AFRL, Syn., UBFC	9.42	14.6	0.436
AFRL, Syn., UBFC, BP4D+	9.37	14.6	0.440

MAE = HR Mean Absolute Error (beats/min), RMSE = HR Root Mean Squared Error (beats/min), ρ = Pearson Correlation in HR estimation.

erated synthetic avatars. For each video, we reduced the resolution of the video to 36x36 pixels to reduce noise and computational requirements while maintaining useful spatial signal. The input to the appearance branch was calculated as the average frame over all T time-points. The input to the motion branch was a set of T normalized difference frames, calculated by subtracting consecutive frames and normalizing by the sum. We used a window size of $T = 30$ video frames to predict the PPG waveform for the corresponding 30 time points. During training, a sliding window of 15 frames was used to increase the total number of training examples. The model was trained for eight epochs using the Adam [7] optimizer, with a learning rate of 0.001.

5. Results and Discussion

A 6th-order Butterworth filter was applied to the model outputs (cut-off frequencies of 0.7 and 4.0 Hz). Standard metrics were computed over all windows of all the test videos in a dataset: mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (ρ) between the estimated HR and the ground truth HR. Table 1 shows the results achieved by our model.

Fig. 2 shows examples of waveforms generated by our model. We do not have access to the ground-truth PPG signals measured via a contact sensor; however, these examples show clear periodic signals, suggesting that the pulse signal was recovered well. There are no other obviously periodic changes in the video, other than respiration and blinking, and these waveforms do not match the frequency and/or dynamics that would be expected from those signals. One way to inspect which regions of the video frames are used by the model to recover the estimated PPG signal is to plot the attention mask weights. Fig. 2 shows attention masks for frames from a subset of the test videos. These examples illustrate that the model correctly learns to segment the participant's face from the background, in addition to highlighting regions of the face containing skin. These regions are crucial for detecting the subtle skin color changes caused by the change in blood flow.

Beat-to-beat (or instantaneous) pulse rate measurement from video remains a challenging task. While many models can achieve average pulse rate measurement accuracy (e.g. over 30 seconds) of close to 1-2 beats/minute mean absolute error, achieving that level of performance on a beat-to-beat level is very difficult. In this work we used

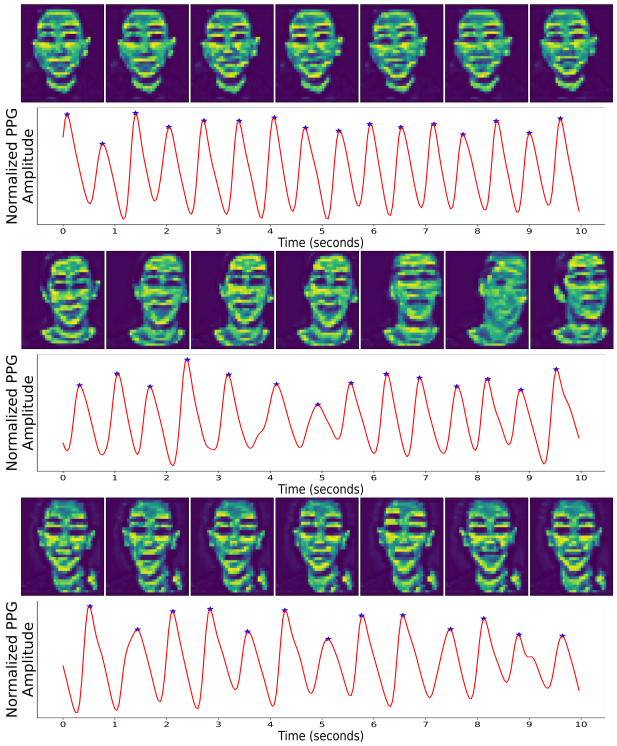

Figure 2. Examples of predicted waveforms and attention mask examples from the V4V test dataset

a model trained on a large corpus of real videos combined with synthetic data. Our final model achieved an MAE of 9.37 beats/minute on the test dataset (see Table 1). Qualitatively, the extracted PPG waveforms had a reasonably high signal-to-noise ratio (see Figure 2). Training with data from a similar distribution (BP4D+) to the testing data (V4V test set) had only a marginal benefit, as the HR MAE decreased from 9.42 to 9.37 beats/minute. The apparent generalization is encouraging. There are other reasons to be optimistic too. Over the past 10 years performance dramatic improvements have been made in many areas of machine learning, including computer vision. Work in camera-based physiological measurement has yet to take advantage of many of these, including unsupervised pretraining and developments in transformer architectures.

6. Conclusion

We have presented a neural architecture, model, and results for camera-based vital sign measurement that captures spatial and temporal information for recovering cardiopulmonary signals from video. To help promote the generalizability of this supervised model, we leverage a set of synthetic avatars during training, alongside real video datasets. Our results show reasonable performance on the V4V challenge data. However, instantaneous (beat-to-beat) pulse rate estimates remain challenging.

References

[1] Vladimir Blazek, Ting Wu, and Dominik Hoelscher. Near-infrared ccd imaging: Possibilities for noninvasive and contactless 2d mapping of dermal venous hemodynamics. In *Optical Diagnostics of Biological Fluids V*, volume 3923, pages 2–9. International Society for Optics and Photonics, 2000.

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

[3] 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.

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

[5] 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.

[6] Goldberger Ary L., Amaral Luis A. N., Glass Leon, Hausdorff Jeffrey M., Ivanov Plamen Ch., Mark Roger G., Mietus Joseph E., Moody George B., Peng Chung-Kang, and Stanley H. Eugene. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, June 2000.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[8] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. *arXiv preprint arXiv:2007.06786*, 2020.

[9] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[10] Daniel McDuff, Javier Hernandez, Erroll Wood, Xin Liu, and Tadas Baltrusaitis. Advancing non-contact vital sign measurement using synthetic avatars. *arXiv preprint arXiv:2010.12949*, 2020.

[11] Rita Meziatisabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021.

[12] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018.

[13] Amruta Pai, Ashok Veeraraghavan, and Ashutosh Sabharwal. Camerahrv: robust measurement of heart rate variability using a camera. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, volume 10501, page 105010S. International Society for Optics and Photonics, 2018.

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

[15] Ambareesh Revanur, Zhihua Li, Umur A. Cifti, 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 Workshops*, 2021.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[17] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007.

[18] 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*, pages 2396–2404, 2016.

[19] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, 2017.

[20] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[21] 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.

[22] Ting Wu, Vladimir Blazek, and Hans Juergen Schmitt. Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes. In *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*, volume 4163, pages 62–70. International Society for Optics and Photonics, 2000.

[23] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 151–160, 2019.

[24] 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.