VitalLens 1.0 Technical Report

Philipp V. Rouast
Rouast Labs
philipp@rouast.com

Abstract

We report the development of VitalLens, an app that estimates vital signs such as heart rate and respiration rate from selfie video in real time. The estimation engine of VitalLens is a computer vision model trained on a diverse dataset of video and gold-standard ground truth physiological sensor data.

1 Introduction

Video of the human face and upper body contains signals with information about the subject's physiological state. Given appropriate circumstances, these signals can be strong enough to estimate vital signs such as heart rate and respiratory rate. This process is known as Remote Photoplethysmography (rPPG); many rPPG approaches have been proposed, including handcrafted algorithms and ones learned from empirical data. Depending on the intended application, we can measure advantages and drawbacks of rPPG approaches in the accuracy, inference speed, and robustness regarding various factors impacting estimation performance.

Paragraph on G, CHROM, and POS.

Paragraph on DeepPhys and MTTS-CAN.

Acknowledge other models we don't compare because lack focus on real-time.

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities and limitations of VitalLens. The estimation engine of VitalLens is a recurrent convolutional model trained on a diverse dataset of video and gold-standard ground truth physiological data.

This report does not contain any further detail about the architecture of the model or training methodology - we reserve this for future publication.

3 Datasets

Two major datasets are involved with training and evaluation of VitalLens: The *PROSIT* dataset and the *Vital Videos* dataset. The vast majority of the data used to train VitalLens comes from PROSIT. We enrich the training data with a subset of data from the Vital Videos dataset, as detailed in this section. The main Vital Videos dataset serves as a test set for the evaluation of VitalLens.

3.1 PROSIT

PROSIT (Physiological Recordings Of Subjects using Imaging Technology) is our in-house dataset for practical rPPG applications.

Participant recruitment and session protocol. As part of our goal of creating a diverse rPPG dataset for practical applications, we recruit participants and collect data at various locations such as residential homes, offices, libraries, and clubs. Each potential participant was required to go through an informed consent process in accordance with Australian privacy laws prior to participation. During the session, participants are asked to complete tasks on an iPad while sensor data is being recorded.

Sensors and collected data. The time-synchronized sensor array used for PROSIT consists of a video camera, electrocardiogram (ECG), pulse oximetry, blood pressure monitor, and an ambient light sensor. This yields a rich set of data including video, ECG, PPG, SpO2, respiration, blood pressure, and ambient luminance. We also collect age, gender, height, and skin type metadata according to the Fitzpatrick scale.

Pre-processing. We pre-process and split each session into small chunks of 5-20 seconds with valid video and signals. As part of this step, we also calculate summary vitals for each chunk from the continuous signals, and extract further metadata such as the amount of participant movement and illumination variation.

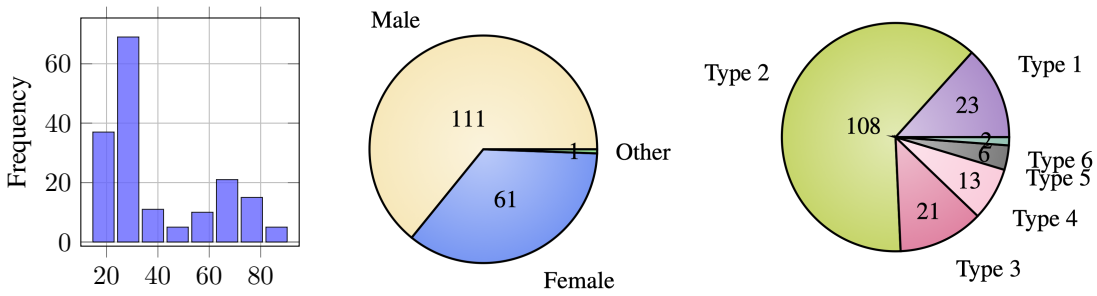
Dataset size and split. Development of PROSIT is ongoing. As of the writing of this report, it comprises 157 unique participants across 173 recording sessions in 45 different locations. This results in a total of 9,767 chunks or 27.8 hours of data.

Table 1: PROSIT Dataset Size

Split	# Participants	# Chunks	Time
Training	114	6,765	19.4 h
Validation	23	1,599	4.5 h
Test	20	1,403	3.9 h
Total	157	9,767	27.8 h

Each participant is randomly assigned to be part of either the *training*, *validation*, or *test* set to ensure that all participants seen during validation and test are previously unseen by the model.

Participant diversity. The demographics of the participants in PROSIT are given in Figure 1. Most participants in PROSIT are under 40 years old, and there are more male than female participants. As we show in Section 5, this is not an issue in practice. However, the skin type diversity of PROSIT is also lacking, with too many type 2 and too few type 5 and 6 participants. This is in fact an issue. To avoid a skin type bias in our model, we supplement our training data with the *Vital Videos Ghana* training set which has mostly type 5 and 6 participants.



(a) Age (b) Gender (c) Skin type

Figure 1: Participant demographics in PROSIT

Vitals diversity. Distributions of the vitals in PROSIT are given in Figure 2. The participants are mostly healthy, so these vitals fall in the typical ranges. There are several participants who have an irregular heartbeat.

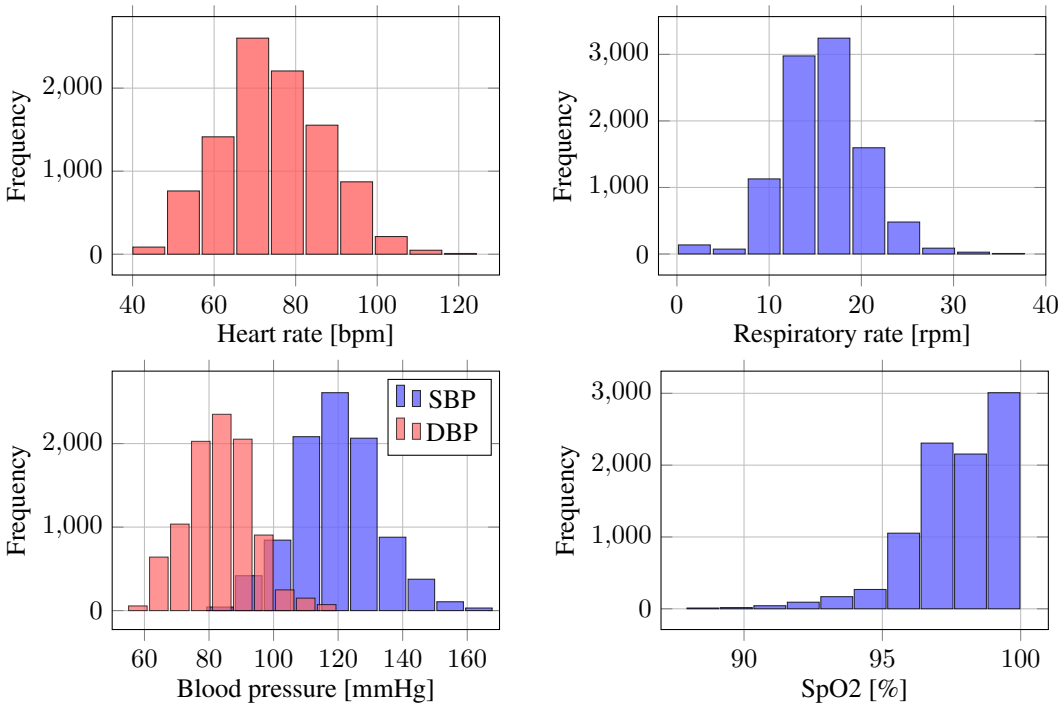


Figure 2: Distributions of chunk summary vitals in PROSIT

3.2 Vital Videos

3.2.1 Vital Videos Medium

Used for test.

Pieter's dataset Number of unique subjects Total duration Statistics Distributions of metadata

3.2.2 Vital Videos Ghana

Used for training.

Pieter's dataset Number of unique subjects Total duration Statistics Distributions of metadata

4 Methodology

?

5 Results and Discussion

5.1 Vitals estimation

Table comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens

5.2 Which factors impact estimation performance?

Regression analysis to determine factors impacting estimation performance

5.3 Impact of subject movement

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR or MAE estimating HR - vs subject movement on x axis. Bar chart comparing DeepPhys, MTTS-CAN, VitalLens SNR or MAE estimating RR - vs subject movement on x axis.

5.4 Impact of illumination variation

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs illumination variation on x axis.

5.5 Impact of subject skin type

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs subject skin type on x axis.

5.6 Impact of subject age

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs subject age on x axis.

6 Conclusion

Authors may wish to optionally include extra information (complete proofs, additional experiments and plots) in the appendix. All such materials should be part of the supplemental material (submitted separately) and should NOT be included in the main submission.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.