PPG Field Study Dataset

I. General information

Contact persons: Attila Reiss, Ina Indlekofer and Philip Schmidt.
Contact at: firstname.lastname@de.bosch.com.
If you publish material based on this dataset, please reference the publication [1].

I.1. Dataset purpose

Photoplethysmography (PPG) is widely used nowadays for continuous heart rate monitoring. However, PPG-based heart rate estimation is still a challenging task, mainly due to motion artefacts. On the other hand, existing publicly available datasets have several shortcomings, especially considering short recordings with a very limited number of activities performed in lab settings (typical datasets include e.g. 5 minutes of data per subject while walking/running on a treadmill). Therefore, our goal is to provide the research community a large dataset with a wider range of activities performed under more natural, close to real-life conditions.

I.2. Dataset usage: quick start

Details about the dataset are given below, in the subsequent sections. However, in case somebody would like to directly "dive in" and use the dataset, we provide a short guide in this section.

The main purpose of the dataset is PPG-based heart rate estimation. For this task, typically three sensor modalities are used: a) the PPG-sensor itself, b) 3D-accelerometer embedded in the same device as the PPG-sensor, used to compensate motion artefacts, and c) ECG which provides heart rate ground truth. It is common practice in related work to use a sliding window approach (window length: 8 seconds, window shift: 2 seconds) [6-10]. This means that all data signal is segmented with this sliding window, and the goal is to determine the heart rate on each 8-second window segment. In order to address this task, the following parts of the dataset are required:

- In each subject's folder, only the file SX.pkl is required, which includes synchronised and labelled data.
- Within SX.pkl, *'data' -> 'label'* includes the heart rate ground truth for the 8-second segments, shifted with 2-seconds (the labels were extracted from the ECG-signal).
- Within SX.pkl, *'data' -> 'signal' -> 'wrist' -> 'BVP'* includes the PPG-signal. The sampling rate is 64 Hz. This signal should be segmented with the same sliding window (8/2 seconds) to get the same time segments as the ground truth is provided for.
- Within SX.pkl, *'data' -> 'signal' -> 'wrist' -> 'ACC'* includes the ACC-signal. The sampling rate is 32 Hz. This signal should be segmented with the same sliding window (8/2 seconds) as the ground truth and the PPG-signal.

I.3. Dataset structure

The dataset is organised so that each subject has a folder (SX, where X = subject ID). Each subject folder contains the following files:

- SX_quest.csv: contains information about the subject; see details below
- SX_activity.csv: contains the activity labels; time is given in seconds and refers to the time when the respective activity started
- SX_RespiBAN.h5: contains data from the RespiBAN device; see details below
- SX_E4.zip: contains data from the Empatica E4 device; see details below
- SX.pkl: contains synchronised data and labels; see details below

I.4. Subjects

15 subjects participated in the study, seven male and eight female subjects, aged 30.60 ± 9.59 years. Information on each subject can be found in SX_quest.csv within the respective subject's folder. An overview is given in this document as well, see Table 1 in the Appendix. The following information is provided there:

- age (years)
- gender
- height (cm)
- weight (kg)
- skin type (according to the Fitzpatrick scale [2])
- fitness level (how often does the subject do sports; on a scale 1-6 where 1 refers to less than once a month and 6 refers to 5-7 times a week)

II. Data collection protocol

Each subject followed a defined data collection protocol, including eight different activities. The duration of these activities was approximately defined as well. However, since the goal of this study was to collect data in a close to daily-life setting, subjects were instructed to carry out the activities as naturally as possible. An overview of the data collection protocol is given in Table 2 in the Appendix. The following list provides details on each of the activities, including the activity ID used in the final data structure (see III.3. below):

- Sitting (ID: 1): Sitting still while reading. The aim of this activity was to generate a motion-artefact-free baseline.
- Ascending and descending stairs (ID: 2): Climbing six floors up and going down again, repeating this twice. This activity was carried out in the main building at our research campus. Note: for subjects S1 and S2, going down was performed only once.
- Table soccer (ID: 3): Playing table soccer, 1 vs. 1 with the supervisor of the data collection.
- Cycling (ID: 4): Performed outdoors, around our research campus, following a defined route of about 2km length with varying road conditions (gravel, paved).
- Driving a car (ID: 5): This activity started at the parking ground of our research campus and was carried out within the area nearby. Subjects followed a defined route which took about 15 minutes to complete. The route included driving on different streets in a small city as well as driving on country roads.
- Lunch break (ID: 6): This activity was carried out at the canteen of our research campus. The activity included queuing and fetching food, eating, and talking at the table.
- Walking (ID: 7): This activity was carried out within the premises of our research campus, walking back from the canteen to the office, with some detour.
- Working (ID: 8): Subjects returned to their desk and worked as if not participating in this study. For each subject, work mainly consisted of working on a computer.

After/before most of the activities, a transient period was included, in order to arrive at the starting location of the next activity (e.g. walk to the stairs at the main building, or walk from the parking ground to the canteen). These transient periods have the activity ID: 0. In total, the data collection protocol took approximately 2.5 hours for each subject. There has only been one major hardware issue, due to which the recorded data of S6 is only valid for the first 1.5 hour.

Figure 1 in the Appendix provides another overview of the data collection protocol, showing the activity labels for S7 together with heart rate information (extracted from the ECG-signal).

III. Data format

Raw sensor data was recorded with two devices: a chest-worn device (RespiBAN Professional, [3]) and a wrist-worn device (Empatica E4, [4]). The activity labels (see Section II above) are synchronised with the RespiBAN raw data (same start time). However, the RespiBAN and the Empatica E4 data need to be manually synchronised. Subjects performed a double tap gesture with their non-dominant hand (where they wore the E4) on their chest. The double tap gesture was performed both at the beginning and at the end of the data collection. The resulting characteristic pattern in the acceleration signal can be used for synchronising the two devices' data. Moreover, the dataset also includes the file SX.pkl, which includes synchronised raw sensor data and labels, see details in Section II.3 below.

III.1. Data from RespiBAN

The RespiBAN Professional was used [3], with the following sensor modalities: ECG, respiration, and three-axis accelerometer. ECG-signal was acquired via a standard three-point ECG. Respiration signal was acquired with an inductive respiration sensor, which is embedded into the RespiBAN chest strap. Three-axis acceleration was acquired via a 3D-accelerometer, which is integrated into the RespiBAN wearable device. All signals were sampled at 700 Hz. Raw data is contained in SX_RespiBAN.h5. Data is organised in a dictionary, corresponding to the sensor modalities. The modalities 'EDA', 'EMG' and 'Temp' include dummy data and should be discarded.

III.2. Data from Empatica E4

The Empatica E4 device was used [4]. The E4 was worn on the subjects' non-dominant wrist. Sampling rate of the different sensors was different, see below. Raw data is contained in SX_E4.zip. When unzipped, the following files contain derived information and thus should be ignored in this dataset: HR.csv, IBI.csv, tags.csv. The file info.txt contains some details on the folder's content. Raw data from the E4 device is contained in the following files (in each file, first line refers to the sensor channel's global timestamp at start, second line refers to the sensor channel's sampling rate):

- ACC.csv: sampled at 32 Hz. The 3 data columns refer to the 3 accelerometer channels. Data is provided in units of 1/64g.
- BVP.csv: sampled at 64 Hz. Data from photoplethysmograph (PPG).
- EDA.csv: sampled at 4 Hz. Data is provided in µS.
- TEMP.csv: sampled at 4 Hz. Data is provided in °C.

III.3. Data synchronisation and labelling

Synchronisation: The double-tap signal pattern was used to manually synchronise the two devices' raw data. Synchronisation was performed in two steps. First, the double-tap at the beginning of the data collection was used to align the start time of the two devices. Second, the double-tap at the end of the data collection was used to correct time drift. This was necessary since, for some of the subjects, the internal clocks of the two devices drifted apart. Therefore, some data samples from the "faster" device were deleted, evenly spaced. However, the required time drift correction was only marginal: The most data to be deleted was for subjects S9 and S10, requiring a deletion of 1.2 seconds out of the over 9000 seconds of the data collection.

Data labelling: This dataset was recorded with the purpose of PPG-based heart rate estimation. Considering this task, reliable ground truth information can be obtained from the ECG-signal. First, an R-peak detector [5] was used. The identified R-peaks were then manually inspected and corrected if required. This was necessary in a few cases for each subject, due to severe motion artefacts on the ECG-signal. Based on the identified and corrected R-peaks, the instantaneous heart rate was computed. Finally, the ECG-signal was segmented with a shifted window approach (window length: 8 seconds, window shift: 2 seconds). Ground truth heart rate is then defined as the mean instantaneous heart rate within each 8-second window. Applying a sliding window with 8/2 seconds is common practice in the literature of PPG-based heart rate estimation [6-10].

The file SX.pkl includes all the above described data, synchronised, and including labels as well. This file is a dictionary, with the following keys:

- 'activity': includes the activity labels, providing IDs 0…8 (see activity list in Section II). These activity labels are based on the file SX_activity.csv. However, in order to make further processing of the dataset more convenient, an activity-signal with 4 Hz sampling rate (which is the lowest sampling rate across all recorded raw sensor data) was created.
- 'label': includes the ground truth heart rate information. As described above, this is provided as the mean of the ECG-based instantaneous heart rate, given on a sliding window of 8 seconds, shifted with 2 seconds.
- 'questionnaire': includes information about the subject, extracted from SX_quest.csv. Details were given above, in Section I.4.
- 'rpeaks': the index of the identified and corrected R-peaks, referring to the ECG-signal. As described above, the identified R-peaks provide the basis of the heart rate ground truth.
- 'signal': includes all the synchronised raw data, in two fields:
    - 'chest': RespiBAN data (all the modalities: ACC, ECG, EDA, EMG, RESP, TEMP). As mentioned above, the modalities 'EDA', 'EMG' and 'Temp' only include dummy data and should thus be ignored.
    - 'wrist': Empatica E4 data (all the modalities: ACC, BVP, EDA, TEMP)
- 'subject': the current subject's ID

Note: as mentioned above, there has been a hardware issue during the data recording of subject S6. The file S6.pkl only includes the valid part of the recorded data, which is about the first 1.5 hour of the data collection protocol.

References

[1] A. Reiss, I. Indlekofer, P. Schmidt and K. V. Laerhoven. 2019. Deep PPG: Large-scale Heart Rate Estimation with Convolutional Neural Networks. MDPI Sensors, 19(14), 2019.

[2] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology 124 (1988), pp. 869–871.

[3] RespiBAN Professional. http://www.biosignalsplux.com/en/respiban-professional (2019). Accessed: 2019-07-15.

[4] Empatica E4 wristband. https://www.empatica.com/research/e4/ (2019). Accessed: 2019-07-15.

[5] P. Hamilton. Open source ECG analysis. In Computers in Cardiology, 2002, ISSN 0276-6547, pp. 101–104.

[6] S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. Chon. 2015. A Novel Time-varying Spectral Filtering Algorithm for Reconstruction of Motion Artifact Corrupted Heart Rate

Signals During Intense Physical Activities using a Wearable Photoplethysmogram Sensor. Sensors 16, 1 (2015).

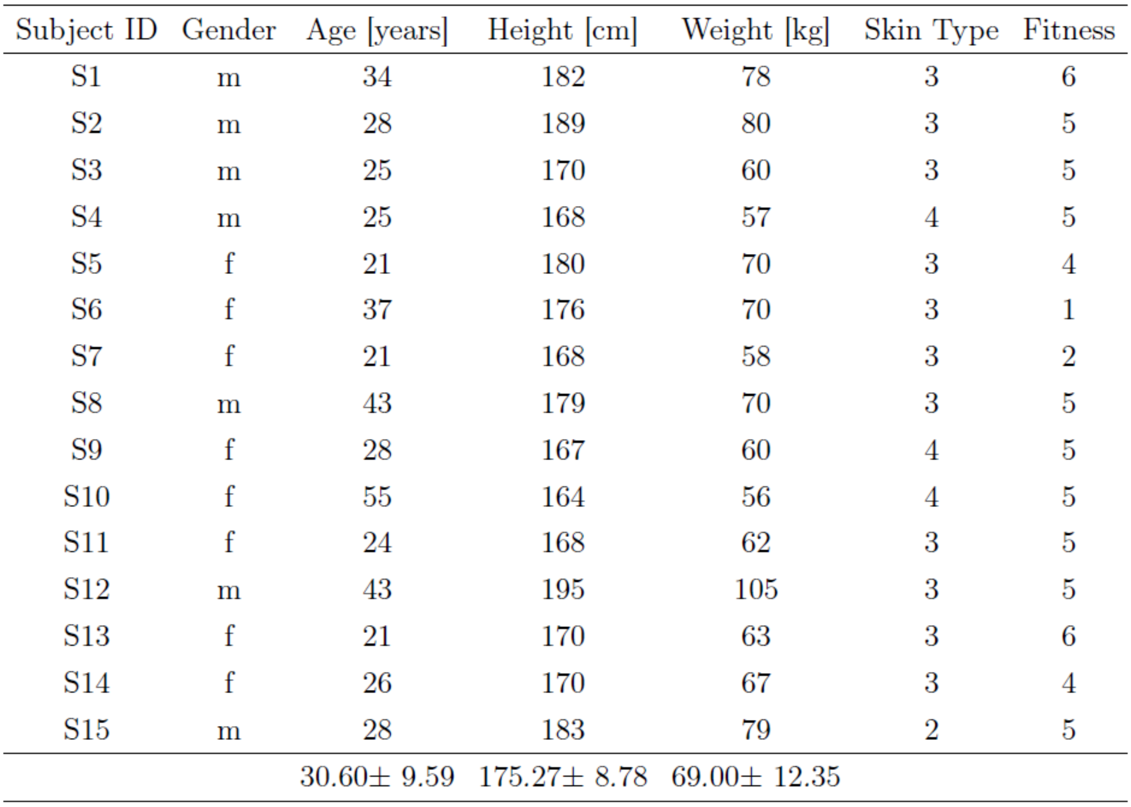
[7] T. Schaeck, M. Muma, and A. M. Zoubir. 2017. Computationally Efficient Heart Rate Estimation During Physical Exercise using Photoplethysmographic Signals. In 25th European Signal Processing Conference (EUSIPCO).

[8] Z. Zhang. 2015. Photoplethysmography-based Heart Rate Monitoring in Physical Activities via Joint Sparse Spectrum Reconstruction. IEEE Trans. Biomed. Eng. 62, 8 (2015), 1902–1910.

[9] Z. Zhang, Z. Pi, and B. Liu. 2015. TROIKA: A General Framework for Heart Rate Monitoring using Wrist-type Photoplethysmographic Signals During Intensive Physical Exercise. IEEE Trans. Biomed. Eng. 62, 2 (2015), 522–531.

[10] A. Reiss, P. Schmidt, I. Indlekofer and K. V. Laerhoven. 2018. PPG-based Heart Rate Estimation with Time-Frequency Spectra: A Deep Learning Approach. UbiComp 2018 Workshop on Combining Physical and Data-Driven Knowledge in Ubiquitous Computing.

Appendix



Subject ID	Gender	Age [years]	Height [cm]	Weight [kg]	Skin Type	Fitness
S1	m	34	182	78	3	6
S2	m	28	189	80	3	5
S3	m	25	170	60	3	5
S4	m	25	168	57	4	5
S5	f	21	180	70	3	4
S6	f	37	176	70	3	1
S7	f	21	168	58	3	2
S8	m	43	179	70	3	5
S9	f	28	167	60	4	5
S10	f	55	164	56	4	5
S11	f	24	168	62	3	5
S12	m	43	195	105	3	5
S13	f	21	170	63	3	6
S14	f	26	170	67	3	4
S15	m	28	183	79	2	5
		30.60± 9.59	175.27± 8.78	69.00± 12.35		

Table 1: Overview of the subjects participating in the data collection.

Activity	Duration [min]
Sitting still	10
Ascending/Descending stairs	5
Table Soccer	5
Cycling	8
Driving car	15
Lunch break	30
Walking	10
Working	20

Table 2: Data collection protocol: activities and their duration.

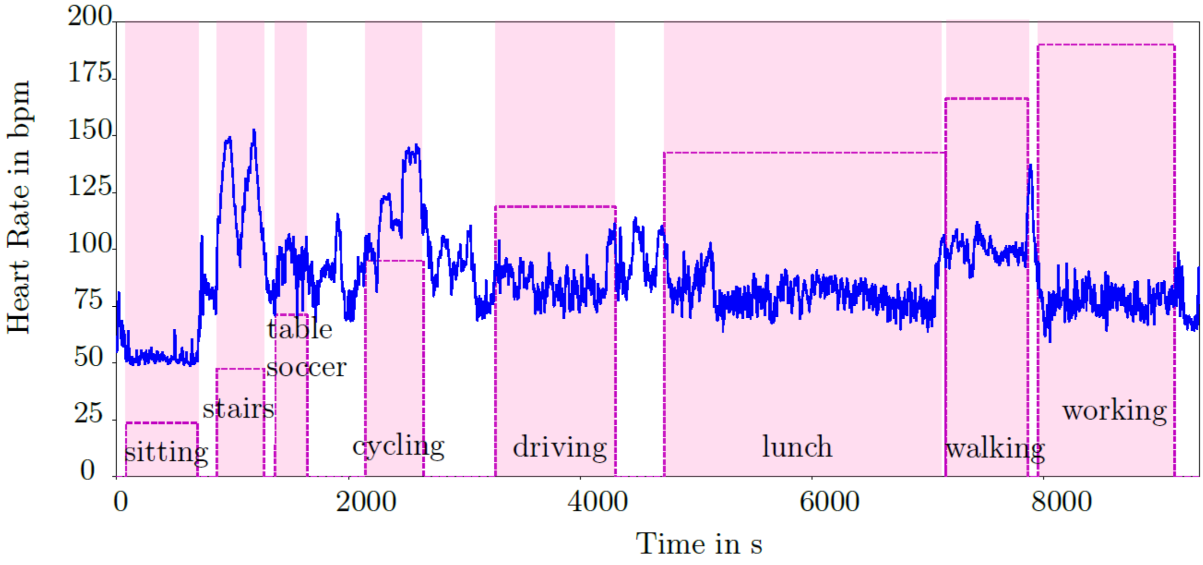


Figure 1: Data collection protocol of S7, including the activity labels and heart rate based on the ECG-signal. The “white” parts between activities refer to the transient periods.