Date of acceptance Grade

Assessor

Accurate and Robust Heart Rate Sensor Calibration on Smartwatches using Deep Learning

Xin Li

Helsinki September 27, 2020

Master's Thesis

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Studieprogram — Study Programme
Faculty of Science	Computer Science

Tekijä — Författare — Author
Xin Li

Työn nimi — Arbetets titel — Title
Accurate and Robust Heart Rate Sensor Calibration on Smartwatches using Deep Learning

Ohjaajat — Handledare — Supervisors
Petteri Nurmi, Eemil Lagerspetz

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	September 27, 2020	77 pages

Tiivistelmä — Referat — Abstract

Heart rate (HR) monitoring has been the foundation of many researches and applications in the field of health care, sports and fitness, and physiology. With the development of affordable non-invasive optical heart rate monitoring technology, continuous monitoring of heart rate and related physiological parameters is increasingly possible. While this allows continuous access to heart rate information, its potential is severely constrained by the inaccuracy of the optical sensor that provides the signal for deriving heart rate information. Among all the factors influencing the sensor performance, hand motion is a particularly significant source of error.

In this thesis, we first quantify the robustness and accuracy of the wearable heart rate monitor under everyday scenario, demonstrating its vulnerability to different kinds of motions. Consequently, we developed DeepHR, a deep learning based calibration technique, to improve the quality of heart rate measurements on smart wearables. DeepHR associates the motion features captured by accelerometer and gyroscope on the wearable with a reference sensor, such as a chest-worn HR monitor. Once pre-trained, DeepHR can be deployed on smart wearables to correct the errors caused by motion. Through rigorous and extensive benchmarks, we demonstrate that DeepHR significantly improves the accuracy and robustness of HR measurements on smart wearables, being superior to standard fully connected deep neural network models. In our evaluation, DeepHR is capable of generalizing across different activities and users, demonstrating that having a general pre-trained and pre-deployed model for various individual users is possible.

ACM Computing Classification System (CCS):
Human-centered computing → Ubiquitous and mobile computing → Ubiquitous and mobile computing design and evaluation methods,
Computing methodologies → Machine learning → Machine learning approaches → Neural networks,
Computing methodologies → Machine learning → Cross-validation

Avainsanat — Nyckelord — Keywords
Heart Rate monitoring, PPG, Performance Evaluation, Deep Learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

Heart rate is the speed of heartbeat measured in beats per minute (bpm) and many factors can cause variation in heart rate of human beings [88]. This makes heart rate a critical indicator of the human physiology. For example, heart rate monitoring has been utilized to detect cardiovascular diseases and abnormality [93]. In the past, heart rate monitoring has been predominantly performed at the hospital or laboratory with dedicated medical devices. The device is usually cumbersome and requires the subject to be connected to the device with cables and electrodes, severely reducing the flexibility of heart rate monitoring in a more ubiquitous scenario. The emergence of the heart rate monitoring belt makes the process portable, however it is inconvenient and uncomfortable for continuous use, especially over a long period.

Continuous heart rate monitoring has become easily available due to the development of wearables that embed non-invasive optical heart rate monitoring sensors on wrist-worn wearables. The continuous access to heart rate information in daily life enables innovative applications and scientific researches in multiple related domains, such as health care, sports and fitness, physiology, psychology, and cognitive science [58]. For example, heart rate monitoring on the wearable has been utilize to assess the intensity of physical exercises [40, 1], to detect chronic diseases [50], to monitor the state of drivers [55, 2], and to infer cognitive or psychological status like emotion and stress [13, 23, 44]. Accurate heart rate measurement is crucial for these emerging applications as all the key information is derived from the heart rate.

Collecting accurate heart rate measurement from a wrist-worn wearable is challenging under certain scenarios, especially where motion is in presence. This is due to the intrinsic characters of monitoring the heart rate on these devices. Current commercial wrist-worn wearables measure the heart rate utilizing *photoplethysmogram* (PPG), which is obtained from *pulse oximeter*. It operates by illuminating the measurement site on the human body with light and measuring the transmitted or reflected light that changes corresponding to the blood circulation [36, 3]. The PPG signal can be utilized to derive heart rate information and other cardiac variables like heart rate variability and oxygen satuation (SpO2). The PPG signal, and consequently the heart rate estimates, are known to be susceptible to *motion artifacts* [77, 99, 59]. Motion artifacts introduce noises in the PPG signal that make the extraction of heart rate information difficult. In addition, motion causes ambient light seeping onto the the photodetector of the PPG sensor to corrupt the heart rate monitoring, especially when the wearable device is not correctly fitted on

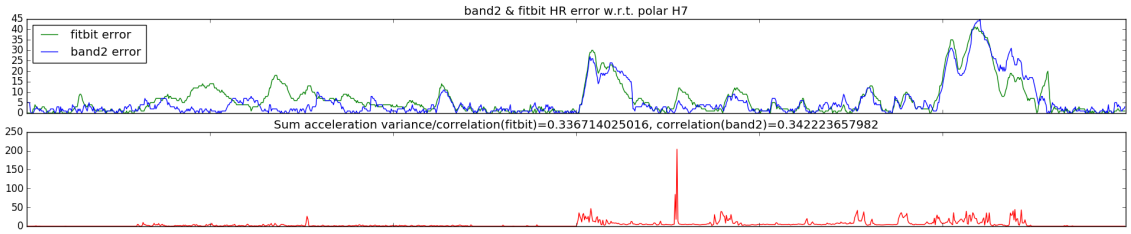


Figure 1: Difference in heart rate measurements (mean absolute error in bpm) between two wrist worn trackers (Microsoft Band2 and Fitbit Surge) and a chest strap HR monitor (Polar H7), and intensity of overall motion as given by the sum of axis-wise accelerometer variances.

the measurement site [97]. More details regarding the PPG signal are introduced in Secton 2.3. Figure 1 illustrate the severity of the issues caused motion artifacts. The difference in heart rate measurements between the two different wrist-worn HR monitors (Microsoft Band2 and Fitbit Surge) and a reference sensor (Polar H7) that is based on electrocardiogram (ECG) is shown in the upper plot, together with intensity of motion (sum of the variance of each three accelerometer axis [7]) depicted in the lower plot. The figure clearly highlights that motion results in increased heart rate error. However, no direct correlation can be observed between the motion and the heart rate errors.

Despite of its vulnerability to motion artifacts, PPG-based heart rate monitoring has been widely integrated into the wrist-worn wearables and used by customers in everyday life due to its pervasive and unobtrusive feature. Previous studies have found good correspondence between the HR measurements on the wearable and the reference HR in rest activities [19, 94], and even in some relatively steady aerobic exercises [82]. This suggests HR measurements from the wearables can approximate the reference HR quite well during rest activities, even being resilient to some trivial motion. However, the performance of the wearables for pervasive HR monitoring in daily life remains unclear when all types of motions are presented.

To better understand the performance of the heart rate monitoring on the wearable in the daily use, in the first part of this thesis we conduct a comprehensive user study to evaluate the off-the-shelf wearables. The experiment protocol in the study comprises of 9 activities that are representative of the human daily activities in terms of the hand, wrist, and body motion present in everyday life. The user study is conducted on 24 participants. We also carry out follow-up studies focusing on isolating specific types of error to further understand HR errors caused by different

factors, such as the contact force and compounded motion. The details of the user study design are discussed in Section 4. According to our result (see Section 5), current heart rate monitoring with PPG sensors is not accurate while different kinds of motions are involved during the measurement. The errors range from 1 bpm to 67 bpm depending on different activities being performed. Also, the variance of the HR error is high within the activity, making it difficult to track the changing trends in heart rate, which is even more important than the heart rate value itself in many applications [62]. The results are analyzed with respect to motion that is quantified by a motion index, suggesting the relationship between motion and HR error is complex and difficult to capture (see Section 5).

In the second part of the thesis, we propose DeepHR as a calibration technique to improve the accuracy of heart rate monitoring on the wearable. The key idea is to associate the motion characteristics captured by an accelerometer and a gyroscope on the wearable with the HR errors that are obtained by calculating the difference between the heart rate measurement of the wearable and a reference sensor (such as a ECG device). However, modeling the relationship between motion and heart rate measurement error cannot be achieved by naive solutions since the heart rate sensors are influenced by multiple sources of motions. As shown by the result of our user study in section 5, both the hand/wrist motion and the body motion can affect the PPG-based HR monitoring. Therefore simply relying on the raw motion measurements is not sufficient to distinguish the motion that degrades the HR monitoring, from the motion that has no significant effect on HR monitoring, for example the imposed motion as the user is riding a car. In addition, modern wearable devices integrate mechanisms for compensate for errors, but the operation of these techniques is not known. These unknown underlying mechanisms make it even more difficult to find the relationship between motion and HR errors. To overcome these challenges, we combined advanced motion sensor representations with deep learning. The sensor representations aim at identifying the sources of motion, while the deep learning model captures the complex relationship between captured motion and the HR errors. The core of DeepHR is a deep learning model comprised of convolutional, recurrent, and MLP layers (see Section 2.5 for preliminary knowledge about deep learning). In deep learning, training data is the data set used for fitting the parameters of the model, to teach the model to learn the relationship and rule described in the data. In the thesis study, training data comprises of motion information and HR errors. It is utilized to teach the deep learning model to learn the relationship between motion characteristics and HR errors. To collect training data,

the user only need to wear both the wearable device and the reference heart rate sensor without any need for manual labeling or recording. Once the DeepHR is trained with enough data, it can be deployed to correct HR errors given the motion information. Actually, a reasonable amount of training data (around 70 hours of data from 10 different users) can enable the model to work with good generalizability (see Section 7).

We validate DeepHR with comprehensive and rigorous benchmarks using data collected in uncontrolled everyday use, and data collected in our controlled user study. With everyday data as validation, the HR error decreases from 10.77 bpm to 6.97 bpm, achieving an improvement of 29.96%, and DeepHR also reduces the variation in the HR error by 33.15%. For the controlled user study data, the improvement ranges from 29.79% to 47.44% as we train the model with different training data. DeepHR is benchmarked against a baseline deep feedforward neural networks that rely on conventional feature engineering technique, demonstrating its superior accuracy and better generalizability across different users and activities (see Section 7).

The contributions of the thesis are summarized here:

- ***Evaluation:*** We evaluate the performance of optical heart rate monitoring on the wearable with 24 participants and 3 different wearable devices. The activities chosen in the the protocol are representative of hand and wrist motion in everyday life. According to our results, the continuous heart rate monitoring on the wearable is not sufficiently accurate in the everyday using scenario. Consequently it is not reliable for the emerging innovative psychological and physiological applications that rely on heart rate information.

- ***Analysis:*** We analyze the user study results with respect to quantified motion to understand the relationship between the heart rate error and the motion characteristics. Our analysis illustrates the accuracy of the heart rate monitoring on the wearable is severely prone to hand and wrist motion and varies considerably across different users and activities. However, the relationship between HR error and motion is complex and cannot be easily captured.

- ***Calibration:*** We develop DeepHR as a calibration model to improve the inaccurate PPG-based heart rate monitoring by learning a function that relates the motion characteristics with the heart rate measurement error. Once the function is learnt, it can be deployed to calibrate the heart rate moni-

toring on a wearable. DeepHR reduces mean absolute error of heart rate estimates by 29.96% in the evaluation with everyday uncontrolled data. With the controlled user study data, the improvement depends on different evaluation settings where the training and validation datasets vary, ranging from 29.79% to 47.44%. DeepHR also reduces the variance of heart rate errors, resulting in a 33.15% reduction in standard deviation of error for everyday data. Additionally, DeepHR offers better generalizability compared to naive deep feedforward neural networks that depend on conventional manually crafted input features.

2 Background

In this section, we first introduce heart rate monitoring in general, followed by detailed introduction on electrocardiogram (ECG) and photoplethysmogram (PPG) that are employed in our study.Then we introduce accelerometer and gyroscope that are used to capture the motion information, and finally discuss the deep learning technique utilized in the thesis to calibrate the heart rate measurement errors.

2.1 Heart Rate Measurement

Heart rate is one of the most crucial parameters of human body and frequently measured to infer the physiological state of the subject. Heart rate can be measured by monitoring different phenomena on human body that are induced by heartbeat and the cardiac cycle. For example, it can be measured based on ballistocardiography by monitoring the subtle motions due to cardiac cycle [35, 91]. These motions are invisible to human but can be monitored by motion sensors and digital cameras. The periodic motion captured by sensors is used to approximate the heart rate. In addition, the variation in blood pressure caused by heartbeat can be used to measure the heart rate. For instance, Kaisti et al. [38] build a system to monitor heart rate based on a pressure sensor. Moreover, heart rate can be measured through the sound produced by heart while beating, known as phonocardiogram [53]. The phonocardiogram signal recorded by microphones can be used to approximate heart rate by applying signal processing techniques [12]. Finally, heart rate measurement is most commonly achieved either by electrocardiogram (ECG) with the electrical sensor that is considered as the clinical golden standard, or by photoplethysmogram

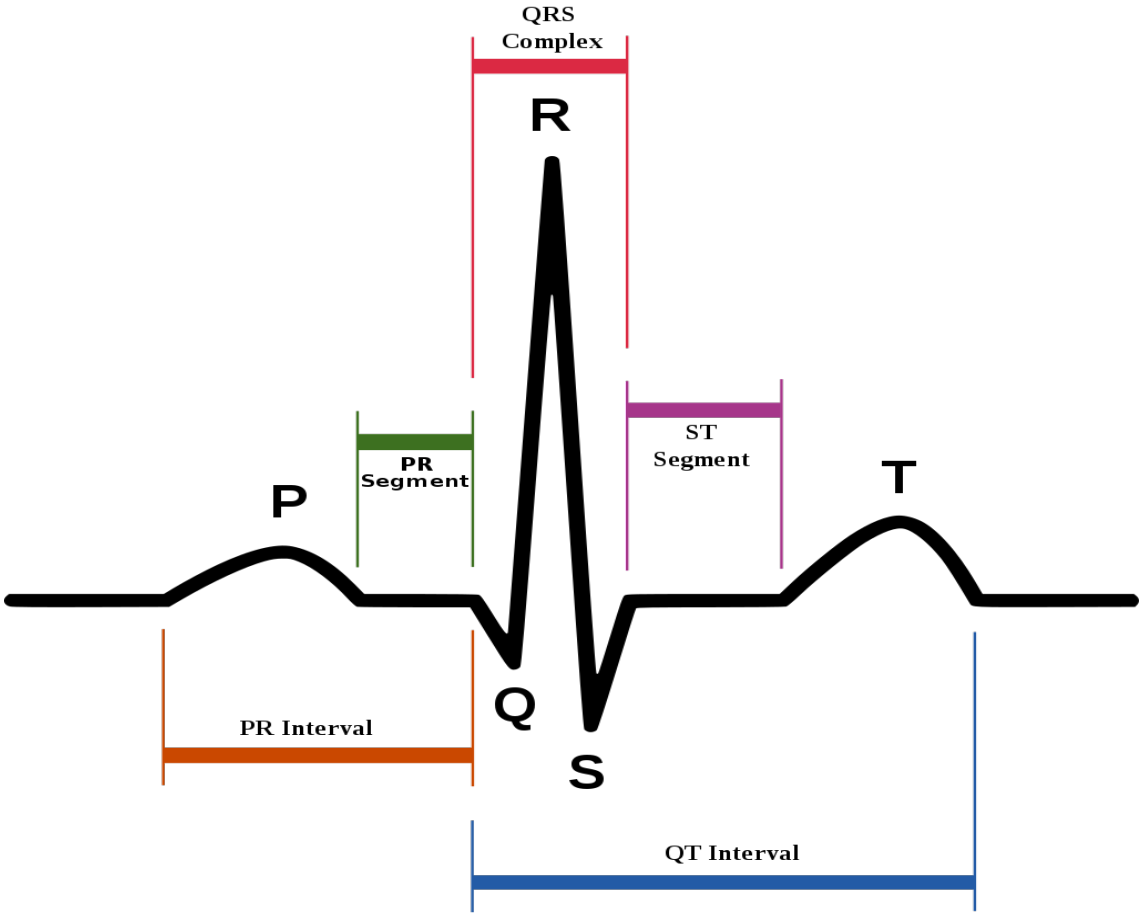


Figure 2: An example of normal ECG.

(PPG) with optical sensor that has been predominantly embedded on commercial wearables. More details of ECG and PPG will be discussed in the following two subsections.

2.2 Electrocardiogram

Electrocardiogram (ECG) is an electrical sensor based technique for heart rate monitoring, widely utilised as golden standard in clinical field for heart rate measurement. The contraction and relaxation of the heart is powered by the electrical impulse during each cardiac cycle. This electrical activity can be measured by attaching electrodes to the skin of the subject. For clinical applications, the electrodes are usually attached to the chest and limbs and connected to a dedicated machine where the collected electrical activity information is processed and displayed. The electrodes can also be integrated into a heart rate monitoring belt to enable a more pervasive

usage of ECG, mostly for sports training and fitness testing. The belt is placed on the chest and the ECG sensor on the belt is connected to other smart devices via Bluetooth for data transmission. Recently, ECG has even being incorporated on Apple watch to indicate irregular heart rhythm with the electrodes built into the crown of the watch [1]. The user need to touch the crown with a finger for 30 seconds to obtain a classification result of the heart rhythm based on the collected ECG signals.

A typical representation[2] of normal ECG signal is shown in Figure 2. It consists of P, Q, R, S, and T waves that represent different phase of the electrical activity of the heart. The electrical impulse is initialized by the sinoatrial node, known as the pacemaker of the heart, which is a specialized structure of the right atrium (the upper part of the heart chamber). Then, the electrical activity spreads through the atria and causes the atria to contract, resulting the blood flowing from atria to ventricles (the lower part the heart chamber). This atrial depolarization that leads to the atrial contraction is marked by the P wave of the ECG signal. The PR interval that begins at the start of P wave and ends at the start of Q wave represents the period when the electrical activity moves from atria to ventricles. The QRS complex [68] comprise of Q wave, R wave, and S wave that appear in rapid and close succession. The QRS complex represents the depolarization and contraction of the ventricles as the electrical activity spreads through the ventricles, which lasts usually between 0.06 to 0.10 seconds for a healthy adult. However, the QRS complex does not necessarily comprise of all the three components because of the possible abnormal conduction of the electrical impulse. For example, a QRS complex can consist of only R wave and S wave while the Q wave is missing. The RR interval is the time period elapsed between two consecutive QRS complexes, which starts at the peak of one R wave and ends at peak of next R wave. The T wave that follows the QRS complex represents the repolarization of the ventricles. Thus, an ECG signal represents a complete cardiac cycle and can be utilized to derive heart rate. Heart rate is mostly commonly derived from ECG by measuring the RR intervals as illustrated in Equation 1.

$$Heart\ Rate = 60\ /\ RR\ Interval\ in\ seconds \tag{1}$$

[1] `https://support.apple.com/en-us/HT208955`

[2] Created by Agateller (Anthony Atkielski), `https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.png`

Many previous studies have utilised ECG based devices to provide heart rate values as ground truth to evaluate PPG based heart rate monitoring devices [80, 56, 90, 82, 14]. ECG is a more accurate technique to monitor heart rate while PPG based heart rate monitoring is more user-friendly. In our study, we chose the ECG-based heart rate monitor Polar H7 as the golden standard to evaluate the performance of PPG-based devices.

2.3 Photoplethysmogram

Photoplethysmogram (PPG) is another alternative technique for heart rate measurements that is based on optical sensors. It makes the affordable and non-invasive heart rate monitoring possible on current smart wearable devices [10]. It measures the variation of blood volume in the tissue and vessel caused by cardiac cycle [36]. The PPG waveform consist of a DC component and an AC component as shown in Figure 3. The DC component is related to the average blood volume and varies slowly according to respiration, while the AC component is closely related to heart rate [3]. In a cardiac cycle, the heart contracts and pumps blood during the systolic phase, and relaxes and fills with blood during the diastolic phase. The systolic and diastolic phases can be captured by the trough and crest of the AC component, which is subsequently used to estimate heart rate. The systolic peaks are detected by applying filtering algorithm to the raw PPG signal [64]. Subsequently, heart rate can be estimated by simply counting the systolic peaks per minute [4], or inferred from the interval between systolic peaks. There are two configuration modes of PPG, *reflective mode* and *transmissive mode*. The reflective PPG illuminates the skin and tissue by a light source and place a photo detector next to the light source to detect the light reflected from the illuminated skin and tissue. It usually utilizes a green light of wave length between 500 and 600 as the light source. The reflective PPG with green light source is currently the most popular configuration as it requires only a single area of contact and can be naturally fitted into daily use by integrating it into the smart watch or band. In contrast, the transmissive PPG detects the amount of light that transmits through the skin and tissue by a photo detector. This requires higher penetration and hence it usually relies on infra-red light of wavelength between 600 and 1300 mm. The measurement site for transmissive PPG is often positioned at the peripheral where the light can penetrate easily [60], like fingers and earlobes. Thanks to its usability and good performance under stationary condition, PPG is widely used for measuring cardiac parameters in both

clinical and everyday use. Next, we discuss factors that affect PPG signal.

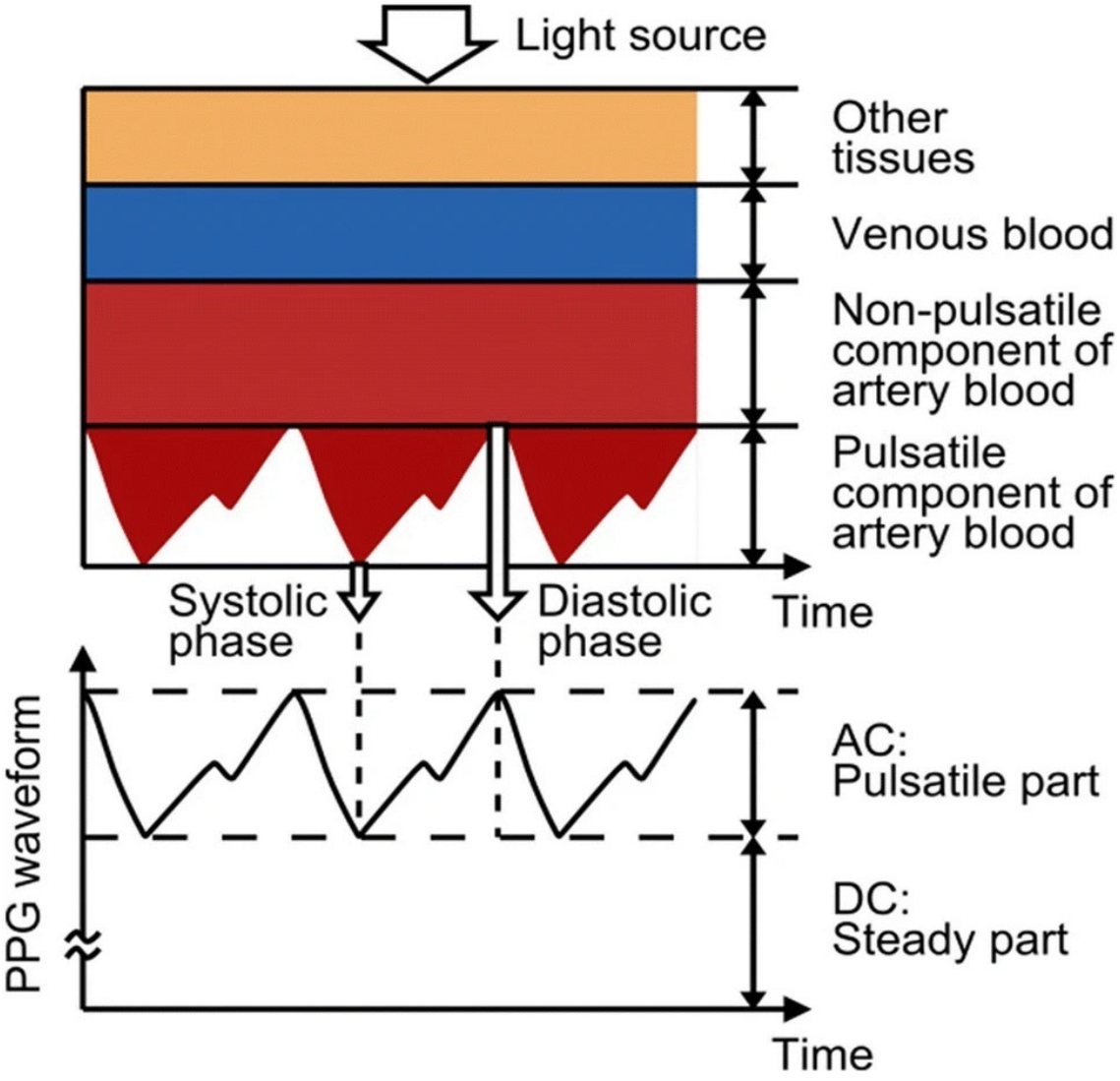


Figure 3: Example of PPG waveform [86].

Motion artifacts in PPG signal caused by motion during the measurement is a major source of error in PPG-based heart rate monitoring. Though the AC component of PPG is essential for estimating heart rate, it only comprises a small portion of the signal amplitude [77]. Therefore, movements that lead to displacement of the sensor and disturb the contact between the sensor and measurement site can easily contaminate the PPG by interfering with the AC component, consequently resulting in inaccurate hear rate measurement [99, 59]. Plenty of research has been focused on motion artifact reduction during PPG measurement, which is discussed in Section 3.2. Besides motion, there are other factors that affect the PPG-based heart rate monitoring. These include skin complexion [20], temperature [60], and

contact pressure [86, 87] (contact force between the sensor and the measurement site), can influence the quality of PPG signal as well.

2.4 Accelerometer and Gyroscope

Accelerometer is a tool used for measuring the acceleration, while gyroscope is a device for measuring the angular velocity or rotating speed. Accelerometer measures the acceleration of an object, the variation in speed with respect to time, by indirectly measuring the acceleration forces, either the continuous static forces like gravity, or the dynamic forces caused by movement. The accelerometers are usually triaxial and the unit is m/s^2. The measurements of accelerometer are sometimes expressed in g, meaning they are relative to gravity. For example, when the accelerometer is placed statically on a horizontal table, the accelerometer measures $-g$ or g inertial force. The gravity is always measured by the accelerometer on earth as it is constantly exerted on all objects. Since gravity is usually stronger than other forces and the orientation of the device may change arbitrarily, it is difficult to measure other forces without eliminating the gravity component first. Gyroscope is a device to measure the rotational motion, the angular velocity on 3 axes called pitch, roll, and yaw, respectively. The unit of gyroscope can be $degrees/s$, rad/s, or $revolutions/s$. Gyroscope is usually integrated with accelerometer on the same chip, because the cost is lower than sum of the individual cost while setting the two separately. Accelerometer and gyroscope have been widely embedded into various devices to collect motion related information for various types of applications, like activity recognition [49, 5], transportation mode detection [33], sports and health [17]. As the motion strongly influences PPG-based heart rate measurements, it is essential to capture the motion related information alongside with the HR measures. In the thesis study, we choose wearables that incorporate both accelerometer and gyroscope to collect the instantaneous raw motion data. It is processed to analyse and characterize the error of HR measurement with respect to motion.

2.5 Deep Learning

Deep learning is a machine learning technique that comprises of multiple neural network layers to learn representations with different levels of abstraction from the data [54]. Deep learning is currently the state-of-the-art technique in many fields, especially in computer vision [89] and natural language processing (NLP) [8]. Besides,

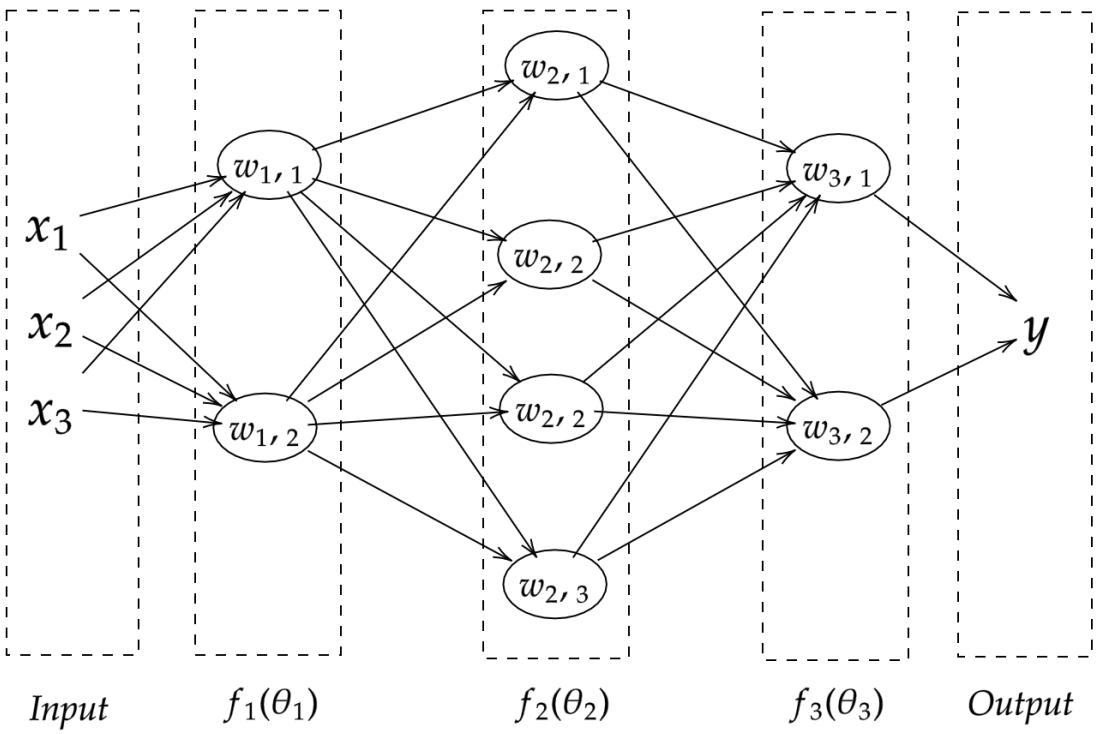


Figure 4: A feedforward neural networks with 3 layers, which accept an input vector of length 3 and output a single scalar, and each layer has 2, 4, and 2 units respectively.

deep learning has been applied to many other fields, like activity recognition [96], mobile sensing [98, 63], and healthcare [67]. In spite of its various application scenarios, deep learning has been based on three kinds of neural network structures: 1) feedforward neural networks, also known as multilayer perceptron (MLP), 2) convolutional neural networks (CNN), and 3) recurrent neural networks (RNN). Feedforward neural network is a fundamental structure in deep learning, and both CNN and RNN can be considered as another special variants of feedforward networks. Depending on the particular applications, with either one or combinations of the three structures, many powerful deep learning models have been constructed. In the thesis study, deep learning techniques are utilised to study the relationship between motion and error of HR measure because of their capability to learn complex patterns from the massive data. With the motion information collected from accelerometer and gyroscope, we apply deep learning to calibrate the HR measurement from wearables. In the following of this section, MLP, CNN, and RNN are briefly introduced together with essential concepts in deep learning.

Feedforward Neural Networks Feedforward neural networks approximate a function f^* that maps the input information x to a target $y = f^*(x)$ [25]. Feedforward neural network forms a function $\hat{y} = f(x;\theta)$ and it learns the optimal parameters θ that minimizes the bias between the approximation $\hat{y}$ and the ground truth value y. The feedforward neural networks usually consists of a number layers. Each layer can be considered as a function that calculates some intermediate results. The layer taking in input x, for example the sensor measurements in our case, is called input layers and the layer outputting the approximation $\hat{y}$ is output layer, whereas the layers in between are called hidden layers. An additional **activation function** can be applied to the output at each layer to enable non-linear transformation. In each layer, the intermediate results of the layer can be calculated using an activation function. Common activation functions are relu, sigmoid, tanh and softmax [71]. Specially, an identity function $f(x) = x$ as activation function means no additional activation is applied to the output. The name feedforward comes from the fact that the information flows from the input x, first to **input layer** followed by the **hidden layers**, where computation happens and intermediate results are produced and transmitted through sequentially, finally to the **output layers** to produce the approximation $\hat{y}$. For instance as shown in Figure 4, a feedforward neural network with 3 layers can be formulated as

$$f(x;\theta) = f_3(f_2(f_1(x;\theta_1);\theta_2);\theta_3) \tag{2}$$

where f_1, f_2, and f_3 represent the three layers, respectively. In this case, f_1 is the input layer (first layer), f_2 is the hidden layer (second layer), and f_3 is the output layer (third layer). The number of the layers is called the **depth** of the model and can be very large, giving the name "**Deep Learning**". **Forward propagation** refers to the process of input x flowing from the input layer through hidden layers finally to the output layer, resulting in a prediction $\hat{y}$ accompanied by a scalar of cost function $J(\theta)$. As in conventional machine learning, a cost function $J(\theta)$ is utilised to evaluate the performance of the model. **Back-propagation** allows the information to flow backward from cost function $J(\theta)$ through the hidden layers to calculate the gradient. With the training data consisting of example pairs $(x, y = f^*(x))$, back-propagation and learning algorithm optimise the weights of the model $f(x)$ recursively to push it closer to the function f^* during the training phase. In many applications the amount of training data and the number of parameters in the model are large. This poses restrictions on model training because of the memory limitation. It is sometimes not feasible and efficient to update the model parameters with all the training data at once. Therefore, the training data is usually split into smaller sets

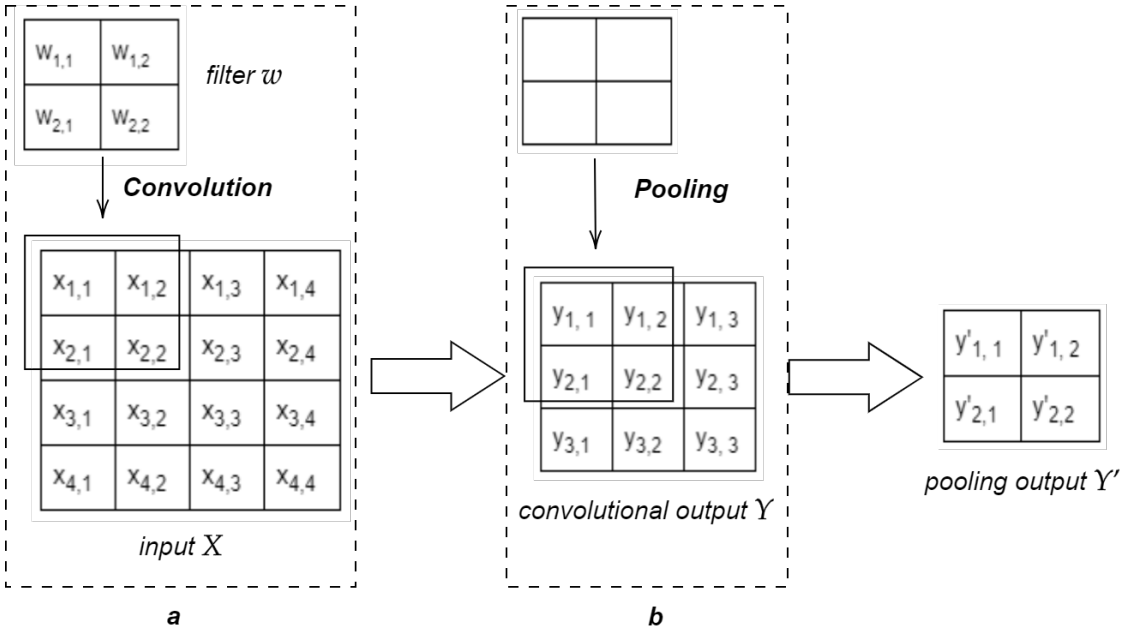


Figure 5: A convolution example with pooling.

called **batches** to perform the back-propagation. The size of a batch is closely linked with the speed and stability of the model convergence. For example, with a small batchsize the model parameters get updated quickly, which may result in the model being far off from the global optima as only limited samples at each batch are utilised for calculation of gradient descent. An iteration of the whole training data set on back-propagation with batches is an **epoch** in training phase, which can lead to either underfitting or overfitting if not set properly. In this thesis, feedforward neural networks comprise part of the proposed DeepHR calibration approach.

Convolutional Neural Networks Convolutional neural networks (CNN) are a specialised structure for processing grid-like data, such as time-series data (1D) and image data (2D) [25]. The layers of a convolutional network apply a **filter** to the input data to perform the convolution operation defined by a **stride** parameter. This is usually followed by a **pooling** operation. As an example of CNN shown in Figure 5, a filter w of 2×2 dimension is applied to the input X of 4×4 dimension to perform the convolution with a stride of 1 at step a, resulting in a 3×4 dimensional output Y, followed by a 2×2 pooling operation at step b, which produces the final output Y'. At step a, the filter w moves over the input Y from left to right and top to bottom, step by step ($stride = 1$) to calculate the convolutional output Y, for example $y_{1,1} = x_{1,1}w_{1,1} + x_{1,2}w_{1,2} + x_{2,1}w_{2,1} + x_{2,2}w_{2,2}$, and $y_{1,2} = x_{1,2}w_{1,1} +$

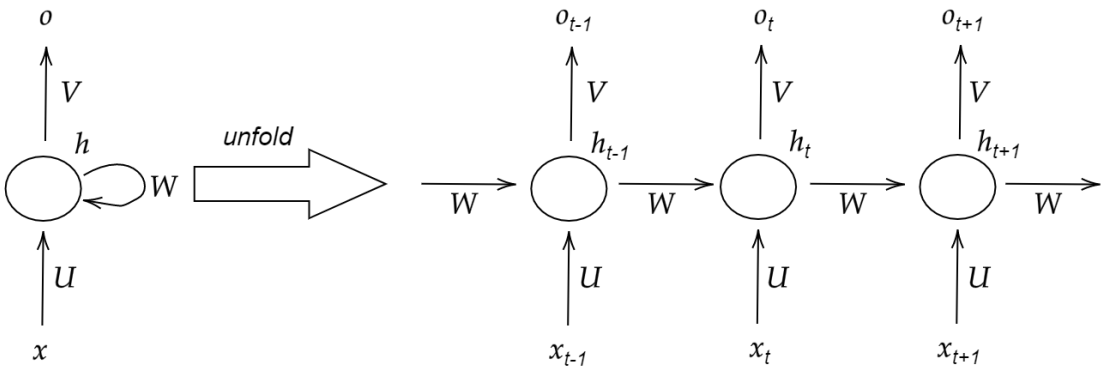


Figure 6: Recurrent neural networks structure

$x_{1,3}w_{1,2} + x_{2,2}w_{2,1} + x_{2,3}w_{2,2}$. Step b illustrates the pooling, either with a max or average function, which usually follows the convolution to further reduce the output dimension, for example using max pooling $y'_{1,1} = max(y_{1,1}, y_{1,2}, y_{2,1}, y_{2,2})$. Compared with feedforward neural networks that multiply the whole weight matrix with all the input as a whole, CNN allows sparse interaction with the data, meaning only a part of the input interacts with the weight matrix (filter) as the filter size is smaller than the input. This sparse interaction significantly reduces the computational overhead, and also reduces storage requirements since the parameters can be shared across operations. Due to its special structure, CNN is effective at extracting features, especially in the field of computer vision like image recognition/classification, and natural language processing. Recently CNN has been increasingly popular in sensor data processing as well [98]. CNN has been applied in the DeepHR approach to process the sensor data collected from wearables for feature extraction.

Recurrent Neural Networks Recurrent Neural Networks (RNN) is another variant of the feedforward neural networks that is specialized for sequence data [25], for example a time sequence x^t, where the time t ranges from 0 to T. As shown in Figure 6, at each time step t, a new value x^t and the hidden state h_{t-1} from last time step are multiplied with the input weight vector $\mathbf{U}$ and the state weight vector $\mathbf{W}$, and an activation function is applied to calculate a new hidden state

$$h_t = tanh(U \cdot x_t + W \cdot h_{t-1}) \tag{3}$$

RNN also produces output at each time step by multiplying the hidden state h_t with

the ouptput weight vector,

$$o_t = V \cdot h_t \tag{4}$$

Depending on the situation, the final output of RNN can be either a single scalar or a vector. Due to the specialised structure of RNN, input at each time step contributes to the final output. Therefore RNN is capable of extracting useful information from the sequence regardless of the positions at which it might appear. However as the the sequence grows larger, RNN suffers from the so-called vanishing gradient problem [37], which stops the model parameters being effectively updated as the gradient becomes very small for the front layers in backpropagation. To mitigate this problem, Long Short-term Memory (LSTM), as a variant of RNN, integrates a gating mechanism to learn long-term dependencies. LSTM is basically a RNN but with better design to pass the states over time steps. Besides the hidden state h_t, there is another cell state C_t going through inside LSTM, which can be considered as the internal memory of the neural network. There are three special gates in a LSTM structure, input gate i, forget gate f, and output gate o,

$$i = sigmoid(U^i \cdot x_t + W^i \cdot h_{t-1}) \tag{5}$$

$$f = sigmoid(U^f \cdot x_t + W^f \cdot h_{t-1}) \tag{6}$$

$$o = sigmoid(U^o \cdot x_t + W^o \cdot h_{t-1}) \tag{7}$$

The three gates are calculated based on the hidden state at last time step and input at current time step in the same way but with different weight vectors **U** and **W**. Input gate decides how much information is allowed to go through from the input at current time step, forget gate decides how much information from previous time step gets through while updating the cell states C_t, and the output gate decides the output from current state. The cell states are updated every time step by first generating a cell states candidate $\tilde{C}_t$ based on the input and hidden state from last recurrence

$$\tilde{C}_t = tanh(U^c \cdot x_t + W^c \cdot h_{t-1}) \tag{8}$$

Then the cell state at current time step are updated with the candidate cell state and the cell state from last recurrence

$$C_t = i \cdot \tilde{C}_t + f \cdot C_{t-1} \tag{9}$$

Then, the hidden state at current time step are calculated based on C_t and the output gate o,

$$h_t = o \cdot tanh(C_t) \tag{10}$$

The output at each time step is obtained by multiplying the hidden state h_t with the output weight vector $\mathbf{V}$, same as in the normal RNN structure

$$o_t = V \cdot h_t \tag{11}$$

In this thesis, LSTM is integrated into the DeepHR to catch the valuable information across the time sequence data to calibrate the noisy heart rate measures.

2.6 Summary

We introduce the preliminary background information of the thesis work in this section. First, the two different types of heart rate monitoring techniques (ECG and PPG) are introduced. The ECG relies on electrical sensors to estimate heart rate, while the PPG relies on optical sensor that allows a more pervasive and unobtrusive way of equipment. However, the PPG-based heart rate measurement is known to suffer from the noise caused by motion and other factors. Then, a brief introduction is given to accelerometer and gyroscope that are utilized to capture the motion artifact. They are known as the inertial measurement unit (IMU) that can be applied to measure the motion information. As we aim to mitigate the motion induced heart rate measurement errors, we choose deep learning model to calibrate the measurement. Before we introduce the details of our model, a brief introduction to the basis of the deep learning is given as preliminary knowledge.

3 Related Work

Heart rate monitoring on wearables has been increasingly popular and also widely studied by researchers. We first review studies on the performance of HR monitoring on the PPG-based wearables. These studies have shown the wearables are capable

of offering accurate estimate of the heart rate during stationary rest activities, such as sitting, standing, and lying. Additionally, some studies demonstrate different levels of the HR monitoring error are observed on wearables for controlled exercises, such as walking, jogging, and running on a treadmill. We introduce and summarize previous studies on the performance of HR monitoring on wearables in this section. Furthermore, a limited number of studies pay attention to evaluate the performance of HR monitoring for non-stationary activities under free-living conditions, which are briefly introduced in this section. Nevertheless, the performance of wearables on HR monitoring remains unclear for daily activities. In our user study, we incorporate 9 different everyday activities to study the performance of HR monitoring under everyday usage scenario. The details of the experiment design are introduced in Section 4. As PPG-based heart rate estimates are susceptible to noises caused by motion, extensive studies have explored approaches of motion artifact removal from the PPG signal to obtain more accurate heart rate. These approaches operates directly on the raw PPG data that is usually unavailable from the wearables. We briefly introduce some of these techniques in this section. Intead of relying on raw PPG signal, we correct the error of heart rate estimate by directly using the actual heart rate values in our study with deep learning technique. Deep learning has been applied for sensing data in many previous studies. We discuss the application of deep learning in sensing field in this section, while our deep learning scheme based on the heart rate and motion sensor data is introduced in Section 6. In this section, we first discuss studies on the performance of PPG-based heart rate monitoring on wearables, followed by introduction to conventional algorithms for correcting the heart rate estimates, ended with discussions on the application of deep learning technique in sensing area.

3.1 Performance of HR Monitoring on Wearables

Studies on HR Monitoring Performance The HR monitoring performance of the PPG-based wearables have been widely studied with various devices. Most of the previous studies on HR monitoring have focused on carefully chosen activities under tightly controlled experiment setups. These studies usually requires the subject to stay stationary [19, 94] or perform under lab conditions during the measurement [82]. Most of these HR rate monitoring devices are capable of providing satisfactory correlation with ground truth heart rate measures during rest activities, and even during some steady-state aerobic exercises these devices offer reasonable

performance. Apart from the rest activities and steady-state exercises, different levels of physical activities, such as standing, walking, jogging, and running, have also been widely studied [41, 56, 90, 92, 84, 18]. Different levels of error have been observed across devices during the physical activities, generally with higher error present when more intense motion is involved. In these studies, all participants are required to perform walking, jogging, and running on treadmill under controlled lab settings. Therefore, whether the reported error patterns would be consistent while the activities are performed freely in daily life remains questionable. There have been a few studies focusing on long-term heart rate monitoring, however most have focused on medical scenarios where motion is strictly limited. For example, Phan et al. [73] tests the performance of heart rate monitoring devices for sleep monitoring purposes and Chudy [14] checks the performance of wrist-worn devices in cognitive tasks. The results of these studies have suggested that the performance of these PPG-based heart rate monitoring devices give satisfactory performance while the motion is low. However, significant variation can be observed as the mean error stays relatively low. As these studies have focused on activities that have very little motion or have simple and highly repeated motion patterns. Therefore, the results from these previous studies are not guaranteed to generalize to everyday scenarios where motions of a wider range and possibly higher intensity are present. This thesis work addresses this issue by evaluating and characterizing the performance of heart rate monitoring on wearable devices, to analyze their performance during everyday activities.

Performance of Wearables in Daily Use Recently there have been studies paying attention to the reliability of wearables under daily usage. Dondzila et al [18] tested the accuracy of step count in free-living situations, but the HR monitoring performance was only validated under lab condition. Reddy et al. [78] assessed HR monitoring performance of two smart wearables with 6 activities of daily living (ADL), suggesting noticeable errors during some daily activities and high variation while the overall bias is relatively reasonable. However, only the overall performance is reported, leaving the details of HR monitoring performance for different types of daily activities unclear. In this thesis, the heart rate monitoring performance under different types of daily activities is assessed and analyzed separately, providing an in-depth understanding of the validity of the PPG-based wrist-type HR monitoring for everyday usage. Consequently the HR measurement errors are characterized with respect to motion pattern, shedding light on the relationship between motion

and error. Specifically, we examine how motions present in common daily activities influence the HR monitoring error, including physical activity, hand motion, and wrist motion. In addition, we discuss how other factors like variation of light signal from optical sensors and strap tightness of the device affect the heart rate estimates.

Table 1 summarizes previous related works discussed in this section on heart rate monitoring performance of PPG-based wearables, including the devices being assessed, the reference devices, activities employed in the evaluation protocol, and the main results.

3.2 Algorithms for Correcting Heart Rate Measurements

Previous researches on correcting the heart rate monitoring have mostly focused on applying algorithms directly on the raw PPG signal to remove errors caused by motion and other sources of noise. The principal idea behind these techniques is to eliminate the noise from PPG signal or decompose the PPG signal into a heart rate component and a motion component. Many techniques have been proposed, ranging from adaptive filters [100, 76], independent component analysis [42], sparse signal decomposition [101] and wavelets [75]. Casson et al. [9] further incorporated motion sensor data from the accelerometer and gyroscope together with the PPG signal to derive motion artifact free heart rate. However, these techniques cannot be applied directly on the heart rate data from consumer-grade wearables, as the raw PPG signal is usually unavailable. In addition, these existing solutions are designed for specific scenarios, like particular sports or intensive activities, under which the motion artifact is easier to be eliminated due to the periodical motion patterns. Therefore these techniques may not be applicable for daily usage scenarios where more subtle and spontaneous motions are present. This thesis work extends the previous studies by developing approaches to calibrate the heart rate directly on the heart rate data without raw PPG information in a pervasive use scenarios.

3.3 Deep Learning for Sensing Data

Deep learning has been widely applied to process various sensing data. In the field of computer vision, deep learning is the most popular solution for tasks like object recognition [32], facial recognition [85], image classification [48]. Deep learning has been integrated into autonomous driving system [27] to overcome some key challenges, such as building perception and reasoning system of an autonomous car.

Study	Devices	Reference	Activities	Participants	Results
[19]	Apple Watch, Motorola Moto 360, Samsung Gear Fit, Samsung Gear 2, Samsung Gear S	Onyx Vantage 9590	rest	4 males, mean age 26.5	accuracy ranged from 99.9% (Apple Watch) to 92.8% (Motorola Moto 360)
[41]	2 Apple Watches on left and right wrists	Polar T13 + Polar S810i	rest, walking, jogging, and running on treadmill	21 males	Correlations (90% CI): walking (L=0.97, R=0.97), jogging (L=0.93, R=0.92), and running (L=0.81, R=0.86)
[56]	Smarthealth wristwatch	ECG	standing, walking, jogging, and running on treadmill	25 participants	valid for standing and treadmill exercise but not consistent when motion is excessive
[90]	Apple Watch, Fitbit Charge HR, Samsung Gear S, and Mio Alpha	ECG	lying, sitting, standing, Walking (Bruce Treadmill protocol), cycling (Ergometer)	22 participants (10 female), mean age 24 (SD = 5.6))	correlation (95% CI): Apple Watch 0.95, Fitbit Charge HR 0.81, Samsung Gear S 0.67, Mio ALPHA 0.87
[92]	Fitbit Charge HR, Apple Watch, Mio Alpha, and Basis Peak, Polar H7	ECG limb leads	walking, jogging, and running on treadmill	50 adults (58% female), mean age 37 (SD = 11.3)	correlation (95% CI): Polar H7 0.99, Apple Watch 0.80, Fitbit Blaze 0.78, TomTom Spark 0.76 and Garmin Forerunner 0.52
[84]	Scosche Rhythm, Mio Alpha, Fitbit Charge HR, Basis Peak, Microsoft Band, and TomTom Runner Cardio	Polar RS400 + WearLink fabric chest transmitter	walking and running on treadmill	50 participants (32 male)	accurate for walking and running, providing high correlation (99% CI) of 0.959, 0.956, 0.954, 0.933, 0.930, 0.929 for TT, BP, RH, MA, MB and FH
[94]	Apple Watch 2, Samsung Gear S3, Jawbone Up3, Fitbit Surge, Huawei Talk Band B3, and Xiaomi Mi Band 2	measured manually	rest	42 participants	MAPE: Samsung Gear S3 (0.04 ± 0.03), Apple Watch 2 (0.07 ± 0.08), Fitbit Surge (0.08 ± 0.12), Xiaomi Mi Band 2(0.12 ± 0.13)
[82]	Apple Watch, Basis Peak, ePulse2, Fitbit Surge, Microsoft Band, MIO Alpha 2, PulseOn, and Samsung Gear S2	ECG	sitting, walking, running, cycling	60 volunteers (29 male), mean age 38 (SD=11)	median error rates range from 1.8% (0.9%-2.7%) at ergometer, to 5.5% (3.9%-7.1%)
[14]	Microsoft Band 2	ECG	N-Back Task (cognitive task)	30 females (mean age 18.67 (SD = 1.69)), 19 males (mean age =21.26 (SD = 4.39))	MSB2 is valid for HR measurement in the selected cognitive task
[73]	LG G Watch R	Pulse Oximeter (CMS-60D), ECG-PowerLab + ADInstruments	rest (10 minutes) and sleeping (4 to 6 hours)	4 participants	reasonable accurate with RMSE of 3.48 bpm (Pulse Oximeter) and 3.54 bpm (ECG Powerlab), correlation $0.89 - 0.90$, showing potential for sleep monitoring application
[18]	Fitbit Charge HR, Mio FUSE	Polar T31	walking, jogging on treadmill	23 female and 17 male	FB shows trend of underestimating the HR, which amplified as HR rises, while MF perform well with mean HR with 1.1 bpm with Polar
[78]	Fitbit Charge 2, GarminvĀvosmart HR+	Polar H7 + Polar A300	standing, walking, and running on treadmill, cycling (ergometer), HIIT, 6 ADLs	20 adults (11 females), mean age 27.5 (SD = 6.0)	reasonably accurate with overall negative bias, $-3.3\%(SD = 16.7\%)$ for Garmin, $-4.7\%(SD = 19.6\%)$ for Fitbit

Table 1: Summary of evaluation on heart rate monitoring devices

Audio sensing is another area where deep learning offers effective solutions. Graves et al. [26] proposes an approach based on RNN for speech recognition, while Lee et al. [57] utilizes deep convolutional networks for audio classification. Deep learning is applied to build systems that are robust to noises for audio sensing tasks [51], such as inferring daily activities (eating, coughing, and driving), detecting the ambient environment, and deducing the user states (stress and emotion). In addition to its application on single-modal sensing data like images and audios, deep learning is effective to combine data from different modalities for content retrieval or human activity recognition [11, 79, 72]. Yao et al. [98] presents DeepSense, a framework to effectively fuse multi-modal sensor input, which can be applied to either regression or classification problems by adapting the output layer of the framework. The CNN structure in DeepSense allows the capability of effectively extracting and fusing the features from multiple sensors, while the RNN structure enables modelling of the temporal relationship, resulting the ability to learn the comprehensive temporal-spatial dependency from the multi-modality sensor data. DeepHR builds upon the foundation of DeepSense, however, it targets on the calibration of heart rate measurement instead of simply object or activity recognition and no applications of deep learning have been found on calibrating the heart rate sensing measurements collected from wearables. We apply deep learning based approach to calibrate the heart rate monitoring on smart wearables, directly utilizing the heart rate together with motion information. Unlike most previous works, our approach works directly on heart rate instead of the raw PPG signal or the RR intervals, without need to design heavily hand-crafted motion features extracted from accelerometer and gyroscope.

3.4 Summary

Previous studies have shown that reasonable accuracy on rest activities or steady-state exercise where the motion involved is either negligible or shows clear patterns. How these results generalize in a more pervasive daily using scenarios is unclear. In the few studies to target the performance in daily activities, the focus is either not on heart rate monitoring or the details of performance for each assessed daily activity is not reported. This thesis work assesses the validity of the PPG-based HR monitoring on wearables in everyday usage, where a wider range of motions are involved. We report and analyze the heart rate monitoring performance under each of the assessed activity as well as characterize the heart rate measurement errors with respect to motion. Though algorithms have been studied to correct the motion

artifact induced heart rate measurement errors, they exclusively focus on making corrections from the raw PPG signal. This thesis presents a deep learning based approach to calibrate the heart rate measurements directly using the heart rate date instead of PPG signals, shedding light on how deep learning techniques can be applied to calibration of heart rate monitoring.

4 Experiment Setup

We evaluate the performance of wrist-worn heart rate monitors through controlled user studies, which consists of a main and a supplementary user study, demonstrating that they are prone to considerable errors resulting from different intensities of motion. To provide robust and accurate estimates of heart rate for sensing applications that aim at inferring physiological and psychological stats of the user, such as overall health condition, emotion, cognitive load, and stress level, it is necessary understand the suitability of consumer grade wrist-worn heart rate monitoring wearables for these applications. We pay special attention to subtle and irregular motion as these are the main source of error in physiological and psychological sensing applications, and as their influence on wrist-worn heart rate monitoring wearables has not been studied much yet.

Our main user study considers 9 activities, covering rest, vigorous activities, and activities involving subtle and irregular motions. The 9 activities are divided into 3 blocks. During the study, the order of the blocks is counterbalanced across participants while the order of activities within a block is kept constant. In total, 24 participants were recruited, internally split into two groups of 12. The first group used a Microsoft Band 2 (MSB2) and a Fitbit (FS) Surge, the second group using Samsung Gear S3 Frontier. We choose both old devices that are nowadays obsolete (MSB [3] and FS [4]), and more modern devices (Samsung Gear S3 Frontier [5]). Incorporating devices of different generations offers the opportunity to see how performance of the wearable has evolved. In the supplementary user study, we conducted follow-up experiments where specific artifacts are specifically controlled and isolated, to further understand the heart rate monitoring performance with finer granularity. In

[3] https://support.microsoft.com/en-us/help/4467073/end-of-support-for-the-microsoft-health-dashboard-applications

[4] https://community.fitbit.com/t5/Surge/Fitbit-Surge-Is-Fitbit-Surge-being-discontinued/td-p/1792210

[5] http://doc.samsungmobile.com/SM-R760/BRI/doc.html

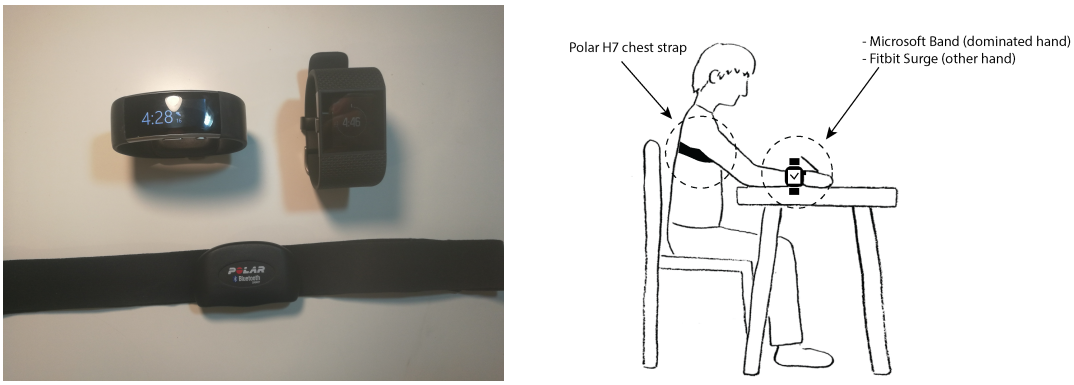
	#	WA	PA	DUR (min)	Task
Block A	1	Med	Low	3	Typing on computer.
	2	High	High	5	Rope jumping.
	3	None	None	3	Lying down on a sofa.
Block B	1	Low	Low	3	Folding clothes.
	2	Med	Med	5	Walking along a predefined route that includes several stairs and doors to open.
	3	None	None	3	Standing still.
Block C	1	High	Low	3	Playing with a Rubik cube.
	2	Med	High	5	Playing a motion controlled game (Saving penalty shoots on Kinect Sports on Xbox 360)
	3	None	None	3	Sitting down in a chair.

Table 2: Description of the experimental tasks. The columns WA and PA correspond to levels of wrist and physical activity. The tasks were performed in blocks of three, where the ordering of tasks was constant within each block, but the order of blocks was counterbalanced across the participants.

the following part of this section, the experiment setups are detailed.

4.1 Main User Study

Participants For the first set of users, who performed the experiment with Microsoft Band 2 and Fitbit Surge, we conducted the study with 12 participants (6 female) consisting of students and faculty staff, who are from different countries in Asia, Europe, South America, and Africa. The median age of the participants was 24 (IQR = 5) and the mean age was 24 (SD=3). For the second group, 12 adults (6 female) were recruited with same standard as in the first set from different countries in Asia, Europe, South America to ensure the diversity of participants as well, with a median age of 28 (IQR = 9) and mean age of 30 (SD = 5). In both groups, participants were healthy adults without any known cardiovascular or pulmonary



(a) Set of devices used for data collection (b) Elaboration on devices wearing.

Figure 7: Device setup for collection of data.

diseases, or rhythm issues. Participants were required to avoid heavy physical exercise and beverages with caffeine for at least two hours before the experiment. Data collection was carried out according to local IRB guidelines and participants signed their written consent for recording and using their data.

Apparatus Heart rate measurements were simultaneously collected from the PPG based wearables and a ECG heart rate monitoring belt used as ground truth for both groups of users in the experiment. For the first group of users, a medium size Microsoft Band 2 (MSB) was placed onto the wrist of the participant's dominant hand, while a Fitbit Surge (FS) was placed on the wrist of the non-dominant hand, and a ECG based Polar H7 heart rate belt was worn on the chest of the participant. The MSB was chosen due to its good programmability[6], access to heart rate data, and availability of suitable motion sensors (accelerometer and gyroscope). The Fitbit Surge was chosen due to its popularity, and add another dimension of comparison. The ECG based Polar H7 was used to provide a reference baseline since evaluations have shown it to have good correspondence with ambulatory heart rate monitors [92, 45]. The MSB was placed on the dominant hand to capture the wrist and hand motion information using the embedded gyroscope and accelerometer of the MSB. Note that putting both sensors together to the dominant hand was inappropriate as it would decrease wearing comfort and signal quality, and cause one of the sensors to have sub-optimal measurement site. As the Fitbit used in the experiment did not have programmable API for accelerometer and gyroscope, the experiment was not

[6]Programmability support for MSB has since been discontinued and recently the whole software support as well

repeated with Fitbit placed on the dominant hand. Note that the focues of the study is on characterizing the performance of heart rate monitoring in everyday situations, rather than comparing the performance between the two devices. The devices are shown in Figure 7a and the way they are equipped in the study is illustrated in Figure 7b. In the second group, only one wrist-type wearable Samsung Gear S3 Frontier were equipped on the participant's dominant hand. From Samsung Gear S3 Frotier, we collect heart rate measurements and the same motion information that were previously collected from Microsoft Band 2. This implemented by developing a data logger application on the watch via the Tizen programming platform[7] provided by Samsung. The amount of green light that is reflected from the user's skin is also available and collected for analyzing the quality of the PPG signal.

Design The experiment consists of three blocks, each containing three tasks. The tasks are designed to cover different levels of physical activities (rest, everyday activity, and small to intermediate/intense physical activity), and wrist and hand motions (small, medium/intermediate, extensive). The tasks are also designed to simulate the possible spontaneous activities that are likely present in everyday scenario. The tasks we considered are detailed in Table 2 and include *Typing*, *Jumping*, *Lying*, *Folding*, *Walking*, *Standing*, *Rubik*, *Gaming*, and *Sitting*. The order of tasks is constant within each block with the first activity corresponding to an activity with hand or wrist motions, second to an activity with relatively higher physical activity, and the third consisting of a rest period. Between each task the participants were asked to rest for at least around half a minute until the influence of previous activity on heart rate waned, and an additional 3-minute break was given in between blocks. Due to the different fitness levels of the participants, the needed recovery time after intense activity like *Jumping* varies, therefore we gave additional break time for participants if needed. The order of blocks was counterbalanced across participants to avoid any possible order effects, with all 6 possible permutations of blocks employed twice for two participants chosen randomly. The duration of each task was chosen to be between $3-5$ minutes to ensure the overall feature of heart rate can be captured, while at the same time keeping the duration of the study reasonable (≈ 50 minutes) for the participants. Both groups of user study followed the same protocol except for *Lying* was replaced by *Sitting* due to chnages in the layout of the facility where the experiment was conducted, resulting in two same sitting activities labelled as

[7] `https://developer.tizen.org/development/guides/native-application/location-and-sensors/device-sensors?langredirect=1`

Sitting1 (after rope jumping) and *Sitting2* (after gaming).

Procedure Before the experiment started, the participant was asked to put on the wrist-type heart rate monitors, MSB and Fitbit Surge (the first group of user), or Samsung Gear S3 Frontier (the second group of user), and ECG-based heart rate monitor Polar H7 on the chest (for both group). The wrist-type wearable was placed on the participant's wrist and the experimenter checked it was properly worn. While the user was suggested to fix the watch strap tightly on the wrist, we allowed the user to adjust it reasonably to guarantee the wearing comfort as our research focus is on how motion influence the heart rate monitoring performance in daily usage, where wearing comfort is a crucial factor. Regarding the effect of tightness on HR monitoring performance, a complementary user study was carried out – this study is discussed in Section 4.2. Each participant was then asked to perform the different blocks in Table 2. All the 6 permutations of the 3 blocks were employed twice for a group of users, resulting in a fixed set of 12 orders, which were randomly chosen for a user. Before starting the experiment, the experimenter explained tasks in the protocol and answered questions regarding the experiment if raised. During the experiment the experimenter accompanied the participant and supervised the process through out the experiment, holding an Android smartphone that was logging data from the MSB and polar H7, while with Samsung watch and Fitbit Surge the data was logged on the watch and retrieved later.

The three activities are arranged within a block such that the participant began with a task of trivial physical intensity but constant hand/wrist motion for three minutes. This was done to simulate spontaneous and irregular motion (*Typing*, *Folding* or *Rubik*). It is followed by a task of higher physical intensity (*Jumping*, *Walking*, or *Gaming*) for five minutes, and finally finished with a three-minute rest task (*Lying*, *Standing*, or *Sitting*). The participant was instructed to perform the first two tasks naturally, as in real life, while during the last rest activity the participant was asked to stay stationary as much as possible. In block A, the participant first typed a specific paragraph of text sitting in a chair, which was followed by rope jumping. The duration of rope jumping was adjusted shorter according to the participant's fitness level and physical state if they could not finish the task in five minutes. This was done to ensure we cover a comprehensive range of heart rate values during the intense activity, but simultaneously also to avoid overexerting the participant. Due to the high physical intensity of rope jumping, the duration of break was prolonged as needed to allow the participant's heart rate to return to a stable and normal level

before starting next rest task of lying(for the first set) or sitting (for the second set). In block B, the participant started folding and unfolding of a jacket repeatedly while standing in front of a table. This task was chosen as representative of tasks that involves significant hand motion and hand pose changes but little physical activity. In the second task of Block B, the participant walked inside the university building with a pre-defined route, which was designed to cover both downstairs and upstairs walking (both two floors), and opening doors. After walking, the participant was asked to stand still for three minutes as the last rest task in this block. Finally, in block C, the participant first interacted with a Rubik's by constantly rotating the cube instead of trying to actually solve it, in order to cover extensive subtle wrist motions. The second task in block C was to play a Kinect motion capture game, a mini-game within Kinect Sports involving saving football penalties using hands and feet. The last rest activity in block C was Sitting in a chair. A short break of around 30 seconds was given between tasks within a block for transition to next task and stabilizing the heart beat. However, if the heart rate of the participant does not return normal within the short break, the break was prolonged. Once a block of tasks had finished, the participant had a longer break of $3-4$ minutes to recover from the previous activities.

In the first group, after finishing the all the 9 tasks, the participant filled a questionnaire for basic demographic information such as gender and age. In the questionnaire, the participant was also required to indicate their skin complexion category by a subjective evaluation from 6 possible categories ranging from fair to dark, including very fair, fair, medium, olive, brown, and black. The motivation for including skin complexity is that melanin of the skin inherently influences the PPG light and subsequently the derived heart rate as it is highly absorbent to light [20]. For this reason we included a measure of skin complexity by asking the participant to indicate his or her skin type from six degrees, ranging from very fair to black. Participants were also asked to rate the wearing comfort and perceived tightness of both wristword devices on a 5 point Likert-scale anchored at 1 = *very uncomfortable* and 5 = *very comfortable* for comfort level, and at 1 = *totally loose* and 5 = *tightly fixed* for tightness. In the second group, we collected otherwise the same information except the subjective assessment was replaced by another complementary user study. We conducted another set of experiments, which is introduced in Section 4.2, to study how the level of strap tightness is related to the heart rate.

Preprocessing The measurements from different sensor sources, such as heart rate, accelerometer, and gyroscope, were aligned with respect to the recorded timestamps and manually segmented according to the different activities. Within a segment, 15-second measurements from the beginning and end of the activity were removed. This removes ambiguity during the transition between activities and ensures the collected HR and motion data can be optimally associated with the desired activity. This eliminates potential interference from contiguous activities/breaks. For rest activities, only the measurements after the heart rate had stabilized were considered. This can mitigate the discrepancy caused by different levels of exertion during the activities (particularly rope jumping) generally attributed to different fitness levels. The sampled measurements were interpolated since the raw sampling rate was not always consistent. In the first group, we used a 1Hz sampling rate for heart rate, whereas 62.5Hz was used for both accelerometer and gyroscope, as the target sampling rate after interpolation. In the second group, the data segment and process is similar to what has been applied for MSB in the first set of main user study. The amount of green light that is reflected from the skin was also available, which was used to analyze the quality of PPG signal. We used the same 1Hz sampling rate for the heart rate but a higher rate of 100Hz was used for accelerometer and gyroscope, and the sampling rate of reflected light intensity was 20Hz. In addition, we observed that Samsung Gear S3 Frontier drops measurements and returns zero or negative heart rate values when it fails to detect the actual heart rate, while MSB returns a constant instead so it has similar mechanism. These dropping period was removed while calculating the error and the percentage of dropping is also illustrated in Section 5

4.2 Complementary User Study

In addition to the main user study, we also conducted additional controlled study to explore two factors that decrease the heart rate monitoring performance: (i) how the body motion in addition to hand motion influence the heart rate measurement performance, and (ii) how contact force between the device and the measurement site influence the quality of PPG signal and consequently the heart rate. We designed two complementary studies to investigate the aforementioned two factors, which are called as *fake walking* and *real walking*.

Apparatus and Participants Polar H7 and Samsung Gear S3 Frontier were equipped on the participant the same way as in the main user study. In total 6 healthy male university staff were employed in the experiment, aged from $25 - 36$, with a mean age of 30 (SD=5).

Design and Procedure The complementary study is designed to learn how compounded motion influences the HR monitoring. This is done by comparing the performance between *fake walking* where only hand motion is present and *real walking* where both hand motion and body motion are present. We also study how strap tightness affect the HR monitoring by applying different levels of strap tightness for both *fake walking* and *real walking*. In *fake walking* participants were asked to stand at a fixed position and swing their arms to simulate arm and hand motion during walking, while in *real walking*, participants walked inside a building following a pre-designed route on a flat surface, with pace varying from slow to brisk, for around four and a half minutes depending on the individual gait. Arm motion is the only motion in *fake walking* while in *real walking* it is compounded with body motion. This allows us to evaluate whether the compounded motion artifacts cause even more severe performance degradation than a single one. Additionally, to investigate the effect of contact force between the sensor and measurement site, the participants were required to repeat both *fake walking* and *real walking* several times with different notches of the strap that result in different level of contact force. Participants first tightened the strap to the tightest they can, from which the notch scale increased gradually until the watch became too loose on the wrist. Due to individual variances and tolerance of firmness of the watch strap, the tightest strap notch varies among participants as well as the loosest strap notch where the heart rate monitoring stops working. With this setup, *fake walking* and *real walking* were repeated with different watch strap notches to illustrate how the performance of the HR monitoring is affected by the fixing level (contact force) of the watch. This enables us to quantify the tightness of the watch strap objectively in contrast to the subjective assessment with the first group in main user study.

5 Results

In this section, we discuss the results of experiment described in previous section by first introducing the overall accuracy of the heart rate monitoring, followed by

Activity	Error: MSB	Error: FS	Error, % Drop: Gear
Typing	6.95 (SD=7.02)	8.63 (SD=10.75)	4.55 (SD=2.74), 0
Jumping	67.35 (SD=21.28)	45.44 (SD=21.17)	22.33 (SD=8.68), 27.63
Lying/Sitting1	2.97 (SD=3.59)	1.67 (SD=2.19)	1.17 (SD=1.23), 0
Folding	15.28 (SD=8.88)	9.75 (SD=8.63)	12.13 (SD=5.57), 0
Walking	25.95 (SD=15.35)	12.41 (SD=10.86)	9.41 (SD=8.55), 7.68
Standing	3.90 (SD=5.14)	8.95 (SD=14.36)	4.33 (SD=5.10), 3.95
Rubik Cube	11.68 (SD=7.54)	8.05 (SD=4.56)	3.71 (SD=2.46), 0
Gaming	31.55 (SD=19.01)	13.96 (SD=11.03)	5.10 (SD=4.51), 0.75
Sitting/Sitting2	2.02 (SD=2.05)	3.09 (SD=6.86)	1.72 (SD=2.14), 2.42
Overall	**18.63 (SD=7.02)**	**12.44 (SD=10.75)**	**7.16 (SD=4.55), 4.60**

Table 3: Mean absolute error (MAE) of the Microsoft Band (MSB), Fitbit Surge (FS), and Samsung Gear S3 Frontier (Gear) heart rate monitors for different activities. The last column, Drop: Gear, refers to the percentage of measurements where the Gear fails to provide any heart rate information. The measurements of Polar H7 were used as ground truth for assessing error. The overall error has been calculated as a macro-average of the activity-specific.

an analysis of the measurement error with respect to motion, strap tightness, and the reflected light. In addition, more detailed participant-wise results are illustrated and analysed, and later we discuss special cases of an outliers, finally ending with a summary.

5.1 Overall Accuracy

We first assess the overall error of the heart rate monitors for different activities, and demonstrate that the wrist-type wearables suffer from motions present in daily activities, as shown in Table 3. Mean absolute error (MAE) between the MSB/Fitbit and the Polar H7 chest strap measures the error. The activity-wise error is first averaged for each individual, resulting in 9 error values, and the macro-averaged errors are calculated across all participants. We also calculate the average drop rate for Gear as it tends to fail to provide any heart rate information during activities. The dropped measurements are identified by negative or zero HR values for Gear. As the results in Table 3 show, the heart rate errors generally are high, with the

average errors being 18.63, 12.44, and 7.16 bpm for MSB, Fitbit, and Gear, respectively. Note that the overall lower error of Gear is partially caused by its dropping mechanism that gives zero or negative when motion artifacts interrupt the normal measurements, which is different from MSB that returns the latest valid value when the device fails. Besides, the errors vary considerably across different activities, with the MAE ranging from 1 bpm to over 65 bpm as the level of motion increases. Generally, activities that involve intermediate to high physical intensity result in the highest errors, such as walking, motion capture game, and rope jumping. The results for Rubik cube and folding clothes demonstrate even motions with relatively low level of physical activity but frequent hand motion contribute to noticeable errors. With Gear, the folding clothes activity results in the second highest error among all activities for that device. For rest activities, the errors are small and the mean error is consistently within $1 - 3$ bpm, which also is the margin of error for the chest strap. Although the errors of MSB are higher than those of Fitbit, this is probably because Fitbit was placed on the non-dominant hand and consequently experienced fewer hand motions than the MSB. The two heart rate monitors have similar performance for activities involving only small amount of hand movement (rest and typing). Observed from results of activities with little or no dropped measurement (typing, gaming, Rubik playing, and rest activities), the error of Gear is lower than the two much older devices (MSB and Fitbit). Additionally, participants rated the wearing comfort of Fitbit to be better than that of MSB (median rating 4.5 for Fitbit and 3.5 for MSB). The newer device Gear shows better overall accuracy and lower errors for almost all the activities compared to the other two devices. However, it still struggles for activities involving high intensity and constant hand motion. For example, the highest level of error also occurred during physical activities like Jumping and Walking, and high level of hand motion during Folding also caused noticeable error. This demonstrates the motion degrade the PPG-based heart rate monitoring performance on wrist-type wearables.

As an example of the errors, Figure 8 shows common failure patterns for two activities, rope jumping and motion capture game. Both wrist-worn sensors fail during physical activity by consistently underestimating the heart rate. However, while motion causes increased error, from the plots it is difficult to observe any direct relationship between motion and heart rate estimation error. In next subsection, we discuss more about the influence of motion on heart rate measurement errors.

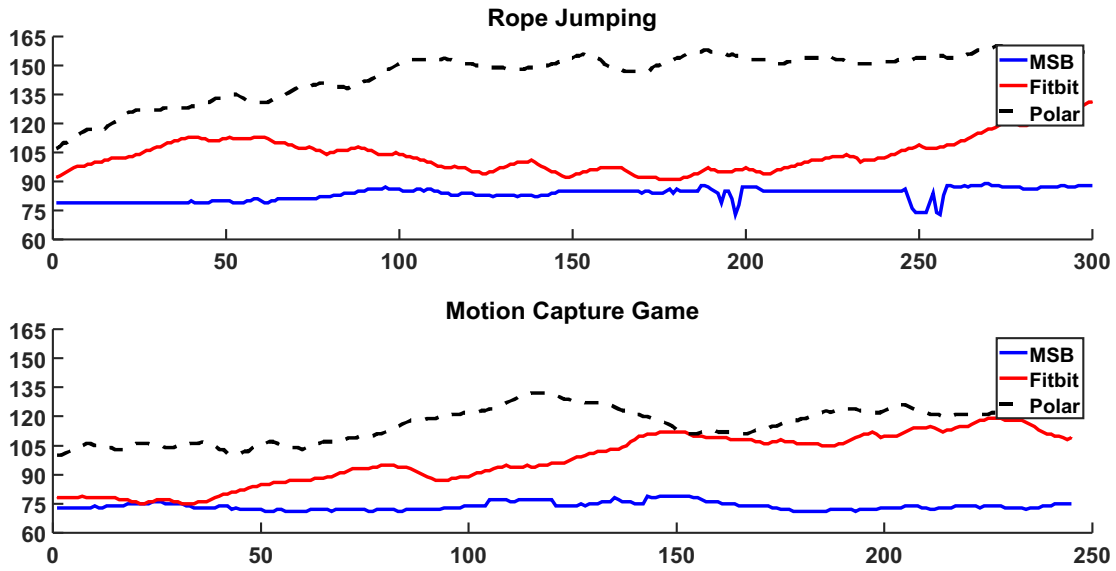


Figure 8: Comparison of heart rate measurements (in bpm) across all three monitoring devices for rope jumping and motion capture game. The plots are created with measurements from one user as a representative example of how the heart rate monitors fail.

5.2 Impact of Motion

In this section, we analyze in more details the relationship between motion and HR error are discussed. To quantify the level of motion, an *motion index* is constructed from the variance of the respective measurements from accelerometer and gyroscope, labeled as acc_var and gyr_var. To construct the activity index, the data is divided into frames of one second with 50% overlap, and the variances of individual axial measurements from accelerometer and gyroscope are calculated within a frame These are then summed up to form the final *motion index* representing the extent of motion in an individual frame. As an example, the sampling rate of accelerometer on Gear is 100Hz, which results in 3 sequences (Seq_x, Seq_y, and Seq_z) of length of 100 corresponding to the triaxial measurements of the sensor consist of a frame. Next, the variance of each sequence is calculated and summed as the motion index acc_var for this frame: $Var(Seq_x)+Var(Seq_y)+Var(Seq_z)$, and the motion indexes gyr_var is calculated in the same way. This approach corresponds to a widely used technique for measuring the level of motion [43, 6]. The motion index acc_var and gyr_var enable quantitative analysis of the motion level, and are used to assess the correlation between the level of motion and HR measurement error.

Activity	Acc. (Corr.)		Gyro. (Corr.)	
	MSB	Gear	MSB	Gear
Typing	0.07 (ρ = -0.05)	0.00 (ρ = 0.03)	0.01 (ρ = -0.02)	0.01 (ρ = 0.02)
Rope Jumping	1.00 (ρ = **-0.32**)	1.00 (ρ = -0.03)	1.00 (ρ = 0.13)	1.00 (ρ = -0.05)
Lying/Sitting1	0.02 (ρ = **0.48**)	0.00 (ρ = 0.10)	0.00 (ρ = -0.10)	0.00 (ρ = 0.09)
Folding Clothes	0.21 (ρ = -0.07)	0.14 (ρ = -0.04)	0.20 (ρ = -0.25)	0.23 (ρ = -0.05)
Indoor Walking	0.30 (ρ = -0.10)	0.08 (ρ = 0.01)	0.18 (ρ = -0.01)	0.23 (ρ = -0.04)
Standing Still	0.18 (ρ = 0.16)	0.00 (ρ = -0.02)	0.00 (ρ = **0.80**)	0.00 (ρ = 0.00)
Rubik Cube	0.00 (ρ = -0.13)	0.03 (ρ = 0.02)	0.04 (ρ = -0.14)	0.05 (ρ = 0.02)
Motion Game	0.75 (ρ = **0.36**)	0.34 (ρ = -0.01)	0.50 (ρ = **0.36**)	0.39 (ρ = 0.00)
Sitting/Sitting2	0.17 (ρ = 0.04)	0.00 (ρ = 0.09)	0.00 (ρ = **0.96**)	0.00 (ρ = 0.10)

Table 4: Normalized level of motion intensity for accelerometer and gyroscope, and Pearson correlation with heart rate error.

Table 4 illustrates the motion index values for the different activities together with the correlations between motion and heart rate error are illustrated. To make the values comparable, the motion indexes across different activities have been normalized by the maximum value of the motion index. Among all activities, rope jumping and motion capture game have the highest levels of motion for both sensors. According to MSB, the correlation between accelerometer derived motion index and HR error is noticeable for both activities, whereas the correlation between gyroscope derived motion index and HR error is only significant for the motion capture game. While intermediate correlations are also observed with MSB for rest activities, there are not significant as both the error and motion levels are low during most of the activity. Even when clear correlations can be observed, they only partially explain the variations in the heart rate error. Accordingly, while motion has a direct relationship with heart rate error, motion information cannot be directly used to compensate for heart rate errors. For the Gear watch, the correlations between HR error and both motion indexes are low, with 0.23 during folding clothes activity being the highest correlation. Note that the lower correlation is likely to result from the Gear dropping measurements during motion, as described in Section 5.1. Additionally, the Gear device is likely to include more advanced motion compensation techniques than the MSB and Fitbit devices, thereby reducing correlations between motion level and heart rate error.

	User	1	2	3	4	5	6	7	8	9	10	11	12	overall
MSB	**Acc Corr.**	0.94	0.87	0.87	0.71	0.94	0.97	0.88	0.91	0.97	0.76	0.96	0.97	0.90
	Gyr Corr.	0.98	0.90	0.97	0.86	0.98	0.97	0.90	0.99	0.94	0.89	0.95	0.97	0.94
Gear	**Acc Corr.**	0.79	0.92	-0.19	0.86	0.37	0.95	0.60	0.60	0.98	0.73	0.60	0.67	0.66
	Gyr Corr.	0.83	0.92	-0.17	0.98	0.42	0.95	0.53	0.82	0.94	0.74	0.55	0.80	0.69

Table 5: Correlation between the mean motion index (*acc_ var* and *gyr_ var*) and mean MAE calculated activity-wise for each participants from study with Gear watch. The MAE and motion index are first averaged over the course of each activities, resulting in two lists of length 9, between which the correlation is calculated.

In Table 5, the activity-wise correlation between mean motion index and mean HR error is calculated to show the strong correlation between HR error and overall motion level over a period of time. With measurements collected from MSB, the overall correlations between error and motion index are 0.9 and 0.94 for *acc_ var* and *gyr_ var* respectively. For the Gear, the correlations are 0.66 for *acc_ var* and 0.69 for *gyr_ var*, much lower than that for MSB, but still reasonably strong, suggesting motion is still a major source error for newer generation of devices. The lower correlation for Gear may be attributed to the dropped measurements discussed in Section 5.1. The inclusion of more sophisticated motion compensation makes it more difficult to detect clear correlation between motion and error in the newer devices. While the instantaneous motion index and HR measurement error are not very closly correlated, the mean motion index of an activity over a period can be indicative of the average HR measurement error during a period.

After analyzing the HR error with respect to overall motion level, we explore more on how compounded motion (body+hand) influence HR monitoring via our complementary study. The detailed result of the complementary study is shown in Table 11 and Table 12. We discuss the effect of compounded motion in this section, while the effect of strap tightness is discussed in Section 5.3. To better understand the effect of compounded motion, we compare the HR error of *fake walking* and the HR error of *real walking* as the strap tightness levels are identical. This is done by comparing MAE in last row in Table 11 and Table 12. For example, at strap tightness level one, the overall MAE of 3.9 in Table 11 is a bit higher than the overall MAE of 3.18 in Table 12. When the strap tightness is adjusted to the normal level (level one to three), lower HR errors are observed in *real walking* without any HR measurement drops. This suggests that compounded motions do not necessarily result in any

User	R squared score			Error
	acc	gyr	acc+gyr	
1	0.45	0.53	0.53	21.74
2	0.42	0.48	0.48	14.84
3	0.16	0.37	0.42	23.62
4	0.31	0.36	0.36	10.17
5	0.38	0.52	0.52	17.91
6	0.44	0.45	0.48	23.54
7	0.23	0.44	0.46	20.49
8	0.34	0.47	0.47	24.34
9	0.24	0.42	0.43	10.79
10	0.34	0.43	0.43	11.30
11	0.51	0.64	0.65	28.99
12	0.20	0.52	0.54	17.38
Overall	0.34	0.47	0.48	18.76

Table 6: R squred score for user study data with MSB

higher HR errors than the non-compounded ones, even when higher HR variation is present in *real walking*.

We also calculated the $R2$ score to analyze the portion of error that can be explained with the motion quantification of different data features, specifically the motion index of accelerometer and gyroscope. To achieve this, three linear regression models are trained based on three different sets of features, *acc_ var*, *gyr_ var*, and the combination of both. The results are shown in Table 6 and 7 for MSB and gear respectively. Overall gyroscope index contributes more to the error for MSB while *acc_ var* illustrates higher impact for Gear measurement HR error. For both device, the combined of gyroscope and accelerometer factors can be used to explain a larger portion of the error.

5.3 Impact of Strap Tightness

The strap tightness (contact force) is another factor affecting the HR measurement errors. As described in Section 4.1, for the first set of main user study, the subjective assessment of strap tightness and wearing comfort information was collected by questionnaire right after the experiment. In Table 8, the collected information

User	R squared score					Error (% Drop)
	acc	gyr	light	acc+gyr	acc+gyr+light	
1	0.16	0.17	0.32	0.17	0.34	6.65 (6)
2	0.20	0.19	0.33	0.21	0.34	4.83 (6)
3	0.02	0.00	0.01	0.03	0.03	13.46 (19)
4	0.15	0.34	0.02	0.36	0.36	9.26 (1)
5	0.15	0.12	0.09	0.15	0.17	9.35 (14)
6	0.45	0.45	0.59	0.46	0.60	8.33 (3)
7	0.07	0.08	0.00	0.08	0.08	1.90 (0)
8	0.07	0.02	0.00	0.08	0.08	1.73 (0)
9	0.45	0.19	0.00	0.48	0.49	5.24 (0)
10	0.11	0.05	0.01	0.11	0.14	4.92 (5)
11	0.09	0.07	0.01	0.09	0.09	6.26 (0)
12	0.20	0.13	0.03	0.20	0.21	8.43 (0)
Overall	0.18	0.15	0.12	0.20	0.25	6.70 (4.50)

Table 7: R squred score for the main user study data with Gear

is summarized. From the result, no significant correlation between error and comfort/tightness can be observed. Participant who gave higher comfort and tightness score are equally likely to encounter high HR measurement errors. However, the subjective assessment standards can be quite inconsistent across different participants, making a reliable analysis of relationship between HR measurement error and strap tightness difficult. The complementary study (Section 4.2) addresses this weakness by analyzing the effect of tightness systematically.

Table 11 and 12 present the motion information together with the measurement error for further analysis of the HR monitoring performance with respect to strap tightness in our complementary user study. As described in Section 4.2, we designed a dedicated experiment to separate effects of compounded motions (body+hand) and non-compounded motions (hand), to study the influence of the strap tightness on the PPG heart rate monitoring performance. This incorporates *fake walking* where the participant only performed hand motion, and *real walking* where the participant performed both hand and body motion. The strap tightness is first set to the tightest level (notch 1) and gradually loosened (to notch 6 or until the watch fails to measure heart rate). Mean absolute error (MAE) and the drop rate are reported in both tables. The average MAE and drop rate are calculated both user-

	Error		Comfort		Tightness	
User	**MSB**	**FS**	**MSB**	**FS**	**MSB**	**FS**
#1	21.74 (5.52)	9.57 (4.78)	5.0	5.0	5.0	5.0
#2	14.84 (5.38)	5.38 (4.01)	3.0	5.0	5.0	5.0
#3	23.62 (6.06)	15.55 (4.55)	2.0	3.0	1.0	2.0
#4	10.17 (4.31)	6.14 (3.09)	4.0	5.0	4.0	5.0
#5	17.91 (5.67)	13.65 (5.25)	5.0	5.0	5.0	5.0
#6	23.54 (4.73)	9.30 (5.60)	2.0	4.0	4.0	5.0
#7	20.49 (5.62)	13.52 (4.78)	4.0	4.0	3.0	4.0
#8	24.34 (7.43)	12.78 (5.45)	1.0	4.0	5.0	4.0
#9	10.79 (4.90)	22.09 (8.60)	2.0	3.0	4.0	3.0
#10	11.30 (4.88)	15.60 (5.45)	5.0	5.0	5.0	4.0
#11	28.99 (5.16)	15.83 (4.25)	5.0	5.0	5.0	4.0
#12	17.38 (6.23)	11.18 (4.11)	2.0	4.0	4.0	4.0
Overall	18.76 (5.50)	12.55 (4.99)	3.3	4.33	4.2	4.2

Table 8: Overall heart rate error for each participant and subjective ratings of comfort and tightness for Microsoft Band (MSB) and Fitbit Surge (FS). Ratings for comfort and tightness were elicited on a 5-point Likert scale with higher values representing more comfort.

User	R squared score								Error (% Drop)
	acc	gyr	light	tight	acc+gyr	acc+gyr +light	acc+gyr +tight	acc+gyr +light+tight	
1	0.03	0.02	0.02	0.40	0.03	0.05	0.40	0.42	7.40 (24)
2	0.04	0.05	0.00	0.19	0.05	0.05	0.20	0.21	9.34 (27)
3	0.10	0.05	0.28	0.34	0.11	0.29	0.37	0.46	11.13 (4)
4	0.03	0.01	0.03	0.03	0.04	0.06	0.05	0.06	11.21 (0)
5	0.20	0.28	0.00	0.39	0.28	0.28	0.41	0.44	8.70 (2)
6	0.00	0.03	0.23	0.45	0.03	0.25	0.45	0.46	8.43 (12)
Overall	0.07	0.07	0.09	0.30	0.09	0.16	0.31	0.34	9.37 (11)

Table 9: R squared score for complementary study fake walking with Gear

wise and level-wise to summarize the overall error. The average MAE is calculated in a weighted manner based on the drop rate as shown in Equation 12.

User	R squared score								Error (% Drop)
	acc	gyr	light	tight	acc+gyr	acc+gyr +light	acc+gyr +tight	acc+gyr +light+tight	
1	0.06	0.04	0.02	0.40	0.06	0.08	0.45	0.46	5.28 (8)
2	0.05	0.09	0.02	0.16	0.10	0.14	0.20	0.34	12.56 (7)
3	0.01	0.01	0.02	0.09	0.01	0.03	0.09	0.10	3.55 (0)
4	0.13	0.35	0.11	0.38	0.36	0.38	0.61	0.61	12.79 (0)
5	0.01	0.07	0.16	0.43	0.07	0.22	0.49	0.49	10.16 (0)
6	0.02	0.06	0.20	0.03	0.07	0.22	0.09	0.22	5.36 (13)
Overall	0.05	0.10	0.09	0.25	0.11	0.18	0.32	0.37	8.28 (4)

Table 10: R squared score for complementary study real walking with Gear

$$MAE_{overall} = \sum_{i=0}^{n} \frac{(1 - DropRate_i) * MAE_i}{\sum_{j=0}^{n}(1 - DropRate_j)} \tag{12}$$

This gives less weight to cases where the MAE is low but drop rate is high. For example, in Table 12, for user 6 at strap tightness level 6, the MAE is 4.15, suggesting the error relatively small, however the drop rate is 80%, which indicates the HR monitoring mostly fails. Equation 12 prevents these cases to lower the overall error. For most users, the error remains low when the watch strap is tightly attached to the wrist, whereas the error and the dropping rate clearly grows higher when the strap is loosened. The loosened strap allows more ambient light leaking onto the photodetecor, and causes the sensor to shift and has more air between the light and skin which reduces light penetration. All these factors contaminates the PPG signal. The looser the watch strap is, the easier the ambient light leak into the PPG light detector, resulting in higher extent of degradation in the HR signal. Fixing the strap tightly mitigates HR errors caused by these factors. We also repeat the $R2$ analysis to study the influence of tightness, with results illustrated in Table 9 and Table 10. The results demonstrate that in both cases the strap tightness is the highest individual error factor found in the complementary user study. However, even in these cases, strap tightness only explains 25-30% of the error. Additionally, the strap levels that resulted in low errors were tighter than what the participants personally chose.

The quality of heart rate sensor light is significantly influenced by the strap tightness and motion artifact. We next use the variance of the reflected green light collected

User	Params	Fake Walking Strap Tightness level 1	level 2	level 3	level 4	level 5	level 6	Overall
	MAE (drop%)	1.13 (0)	2.74 (0)	2.88 (0)	15.94 (27)	19.46 (39)	15.20 (76)	7.40 (24)
	ME	-0.86	1.98	2.13	15.69	18.57	14.66	6.45
User 1	light_var	0.65	1.22	0.40	3.19	349.02	520.21	145.78
	acc_var	6.01	6.28	7.22	8.74	7.98	6.55	7.13
	gyr_var	3.11	3.42	3.49	4.00	3.68	3.26	3.49
	MAE (drop%)	5.13 (0)	5.93 (0)	7.53 (37)	5.60 (0)	28.97 (34)	1.11 (94)	9.34 (27)
	ME	4.06	4.03	7.22	4.30	28.80	1.00	8.29
User 2	light_var	1.33	2.83	364.64	132.20	60.67	176.21	122.98
	acc_var	0.91	1.42	1.38	2.62	3.26	5.81	2.57
	gyr_var	1.09	1.43	1.44	2.09	2.42	3.29	1.96
	MAE (drop%)	4.67 (0)	3.43 (0)	9.54 (0)	10.46 (0)	6.85 (0)	38.51 (25)	11.13 (4)
	ME	2.39	1.66	5.10	10.45	6.00	38.32	9.48
User 3	light_var	3.20	2.45	1.12	0.75	4.12	56.60	11.37
	acc_var	7.68	3.59	4.07	5.88	7.58	8.70	6.25
	gyr_var	4.55	2.60	2.80	3.60	4.28	4.61	3.74
	MAE (drop%)	10.15 (0)	9.60 (0)	13.43 (0)	10.33 (0)	12.56 (1)	N/A	11.21 (0)
	ME	8.85	8.21	11.10	-5.83	7.74	N/A	6.01
User 4	light_var	0.45	0.10	0.06	108.92	579.98	N/A	137.90
	acc_var	2.00	1.22	1.31	3.07	2.23	N/A	1.97
	gyr_var	0.81	0.75	0.67	1.07	0.97	N/A	0.85
	MAE (drop%)	1.30 (0)	5.76 (0)	15.66 (6)	12.51 (0)	N/A	N/A	8.70 (2)
	ME	-0.16	5.27	14.71	12.50	N/A	N/A	7.97
User 5	light_var	1.66	0.94	0.47	6.01	N/A	N/A	2.27
	acc_var	4.57	2.17	1.49	0.99	N/A	N/A	2.30
	gyr_var	2.68	1.76	1.23	0.99	N/A	N/A	1.66
	MAE (drop%)	1.04 (0)	2.18 (0)	3.83 (0)	9.78 (0)	23.38 (31)	19.01 (39)	8.43 (12)
	ME	-0.37	0.50	2.97	9.08	22.60	18.95	7.44
User 6	light_var	0.27	0.14	0.28	1.00	120.68	378.96	83.55
	acc_var	3.13	3.42	4.80	3.87	3.80	3.59	3.77
	gyr_var	2.08	2.37	2.90	2.69	2.71	2.59	2.56
	MAE (drop%)	3.90 (0)	4.94 (0)	8.82 (7)	10.53 (4)	16.82 (21)	26.61 (58)	
	ME	2.32	3.61	7.21	7.70	16.74	18.24	
Overall	light_var	1.26	1.28	61.16	42.01	222.89	282.99	
	acc_var	4.05	3.02	3.38	4.19	4.97	6.16	
	gyr_var	2.38	2.05	2.09	2.41	2.81	3.44	

Table 11: Fake walking.

from the HR sensor [8] to evaluation the quality of the light. The green light value

[8] `https://developer.tizen.org/ko/development/guides/native-application/location-and-sensors/device-sensors?langredirect=1#hrm_green`

User	Params	Real Walking Strap Tightness level 1	level 2	level 3	level 4	level 5	level 6	Overall
	MAE (drop%)	1.44 (0)	1.60 (0)	1.32 (0)	2.47 (0)	7.03 (30)	21.21 (17)	5.32 (8)
	ME	-0.84	-0.37	-0.41	1.18	7.02	19.50	3.75
User 1	light_var	1.40	1.23	0.17	0.69	275.19	39.63	53.05
	acc_var	2.06	2.28	2.75	2.26	2.71	3.54	2.60
	gyr_var	1.67	1.59	1.94	1.62	1.86	2.23	1.82
	MAE (drop%)	3.17 (0)	5.49 (0)	4.55 (0)	42.80 (0)	3.90 (33)	N/A	12.56 (7)
	ME	-0.94	1.39	1.73	42.66	-0.20	N/A	9.58
User 2	light_var	0.39	154.18	77.09	49.48	805.04	N/A	217.23
	acc_var	2.25	3.21	3.54	3.61	3.89	N/A	3.30
	gyr_var	0.76	0.91	1.32	1.30	1.47	N/A	1.15
	MAE (drop%)	2.24 (0)	2.58 (0)	4.02 (0)	2.32 (0)	4.71 (0)	5.43 (0)	3.55 (0)
	ME	-1.07	-1.11	0.06	-0.06	-1.50	4.85	0.19
User 3	light_var	1.95	2.39	3.87	0.55	3.26	7.75	3.29
	acc_var	3.15	4.57	5.04	4.56	6.11	5.34	4.79
	gyr_var	1.99	2.69	2.15	2.40	3.62	3.23	2.68
	MAE (drop%)	8.05 (0)	2.20 (0)	12.96 (0)	18.68 (0)	22.09 (0)	N/A	12.79 (0)
	ME	7.14	0.27	12.63	18.68	22.09	N/A	12.16
User 4	light_var	1.19	0.09	0.22	133.77	64.94	N/A	40.04
	acc_var	2.25	1.74	2.25	2.50	2.86	N/A	2.32
	gyr_var	1.48	1.03	1.40	1.73	1.94	N/A	1.52
	MAE (drop%)	2.14 (0)	7.29 (0)	11.44 (0)	19.77 (0)	N/A	N/A	10.16 (0)
	ME	-0.12	6.18	10.40	19.45	N/A	N/A	8.98
User 5	light_var	1.05	2.44	2.06	80.56	N/A	N/A	21.53
	acc_var	2.03	2.92	2.15	1.98	N/A	N/A	2.27
	gyr_var	1.71	1.99	1.94	1.81	N/A	N/A	1.87
	MAE (drop%)	2.06 (0)	2.21 (0)	1.78 (0)	16.17 (0)	4.79 (0)	4.15 (80)	5.36 (13)
	ME	-0.91	-0.95	-0.55	15.99	-2.12	-2.41	2.12
User 6	light_var	0.40	0.14	0.13	45.31	12.66	53.86	18.75
	acc_var	2.62	2.79	2.69	3.17	3.13	4.13	3.09
	gyr_var	1.82	1.76	1.87	2.37	2.28	2.68	2.13
	MAE (drop%)	3.18 (0)	3.56 (0)	6.01 (0)	17.03 (0)	8.95 (13)	11.78 (32)	
	ME	0.54	0.90	3.98	16.32	5.06	7.31	
Overall	light_var	1.06	26.74	13.92	51.73	232.22	33.75	
	acc_var	2.39	2.92	3.07	3.01	3.74	4.34	
	gyr_var	1.57	1.66	1.77	1.87	2.23	2.72	

Table 12: Real walking.

was first divided by $10,000$ before calculating the variance so that the variance remains within a reasonable range. The light variance index *light_ var* is calculated

User	1	2	3	4	5	6	7	8	9	10	11	12	overall
Light Corr.	0.74	0.98	-0.23	0.37	0.32	0.98	0.76	0.85	0.11	0.75	0.68	0.73	0.59

Table 13: Correlation between the mean *light_var* and mean MAE calculated activity-wise for each participant. The correlation between the MAE and *light_var* is calculated for each activity in the second user study (as the first one did not include light information not provided by MSB). The table shows the average correlation across the 9 activities.

by taking the variance of the data sequence of light signal within a one-second sliding window with 50% overlapping. The light variance index is a good indicator of the light quality, which increase as more ambient light leaks on to the sensor due to looser strap. From Table 11 and 12, we can observe that the light variance increases dramatically as expected once the strap loosened. From the last row of Table 11 and 12, the error and drop rate correlates with the light variance. When the light variance increases, the error and drop rate increases. Besides, we assess the correlation between average MAE and light variance over a period for each individual participant using the data collected from the second group in main user study. In Table 13, the correlation between averaged MAE and light variance stays intermediate to high for most of the users. To calculate the correlation for a single user, the activity-wise macro-average of the HR error and light variance are first calculated, resulting in nine pairs error and light variance, from which the Pearson correlation is calculated. Overall, the loose straps increase the light variance which in turn usually results in higher error.

5.4 Participant-Wise and Activity-Wise Analysis

Heart rate measurement errors varies significantly across participants and activities as shown in Table 8 for MSB and in Table 14 for Gear. The smallest participant-wise error for MSB is around 10 bpm, while the highest error is near 30 bpm. The HR measurement errors for Fitbit are slightly lower than MSB, but similarly are over 10 bpm for most participants. As shown in Table 8, the variation (standard deviation shown in parenthesis) of the HR errors also changes across different participants for the first group of users for the main user study, ranging from 4.31 to 7.43 for MSB, and ranging from 3.09 to 8.62 for Fitbit Surge. In the following paragraphs, we discuss more details on the activity-wise analysis and pay attention to the participant-wise analysis by repeating the same previous analysis but for

Gear user Study											
User	Params	Activity									Overall
		Typing	Jumping	Sitting1	Folding	Walking	Standing	Rubik	Gaming	Sitting2	
#1, male	MAE (drop%)	1.57 (0)	22.35 (40)	0.94 (0)	11.60 (0)	15.30 (10)	1.49 (0)	4.71 (0)	7.96 (0)	1.08 (0)	6.65 (6)
Latin	light_var	1.75	789.77	0.71	29.47	5.87	0.03	5.34	16.16	0.50	94.40
-American	acc_var	0.24	52.95	0.00	10.33	3.41	0.00	2.53	24.95	0.01	10.49
35-40	gyr_var	0.02	9.67	0.00	2.62	0.89	0.00	0.49	4.13	0.00	1.98
#2, male	MAE (drop%)	5.01 (0)	33.70 (58)	1.23 (0)	5.32 (0)	6.26 (0)	2.84 (0)	1.30 (0)	2.61 (0)	2.05 (0)	4.83 (6)
Latin	light_var	2.57	1119.45	0.80	33.76	38.96	0.20	17.27	125.20	0.44	148.74
American	acc_var	0.31	51.91	0.04	13.72	4.56	0.00	2.80	19.45	0.00	10.31
30-35	gyr_var	0.03	10.12	0.01	2.58	1.47	0.00	0.73	4.00	0.00	2.10
#3, male	MAE (drop%)	22.82 (0)	8.76 (86)	2.89 (0)	18.18 (0)	20.15 (9)	34.28 (47)	7.28 (0)	7.66 (0)	2.32 (29)	13.46 (19)
Nordic-	light_var	5.65	623.88	7.20	21.36	17.15	0.17	32.71	227.72	0.32	104.02
European	acc_var	0.45	83.88	0.34	10.44	3.65	0.01	2.69	11.10	0.04	12.51
25-30	gyr_var	0.05	10.29	0.03	1.66	1.36	0.00	0.62	1.38	0.00	1.71
#4, female	MAE (drop%)	1.56 (0)	40.49 (0)	0.70 (0)	14.07 (0)	14.12 (0)	1.48 (0)	1.64 (0)	8.31 (9)	0.86 (0)	9.26 (1)
Western-	light_var	1.51	281.09	12.37	113.42	15.89	0.12	8.03	874.36	0.17	145.22
European	acc_var	0.48	37.42	0.07	12.34	5.50	0.01	1.79	27.08	0.00	9.41
20-25	gyr_var	0.09	9.57	0.01	2.59	3.53	0.00	0.41	3.48	0.00	2.19
#5, male	MAE (drop%)	3.75 (0)	17.80 (59)	2.09 (0)	32.67 (0)	9.34 (70)	1.71 (0)	9.44 (0)	7.71 (0)	4.62 (0)	9.35 (14)
Nordic-	light_var	1.20	441.26	1.12	16.27	22.81	0.31	6.65	12.70	0.08	55.82
European	acc_var	0.19	91.36	0.15	10.68	4.20	0.00	2.02	9.63	0.00	13.14
35-40	gyr_var	0.01	10.03	0.01	1.62	1.37	0.00	0.29	1.30	0.00	1.63
#6, female	MAE (drop%)	8.54 (0)	54.61 (30)	0.80 (0)	8.65 (0)	9.46 (0)	1.51 (0)	1.13 (0)	2.93 (0)	1.40 (0)	8.33 (3)
Nordic-	light_var	41.36	1213.76	2.18	48.73	1.01	0.03	1.16	30.59	0.28	148.79
European	acc_var	0.46	31.14	0.00	8.67	4.81	0.01	1.80	9.77	0.01	6.30
20-25	gyr_var	0.07	6.68	0.00	2.26	1.37	0.00	0.39	1.81	0.00	1.40
#7, female	MAE (drop%)	2.12 (0)	2.35 (0)	0.81 (0)	3.02 (0)	1.52 (0)	1.49 (0)	0.81 (0)	3.16 (0)	1.78 (0)	1.90 (0)
Nordic-	light_var	5.13	1.05	0.25	5.77	2.86	0.04	0.86	5.12	0.26	2.37
European	acc_var	0.52	47.14	0.09	12.19	3.73	0.00	1.72	36.95	0.00	11.37
35-40	gyr_var	0.05	10.08	0.01	2.40	1.17	0.00	0.27	4.82	0.00	2.09
#8, female	MAE (drop%)	1.06 (0)	2.52 (0)	0.84 (0)	2.48 (0)	2.24 (0)	1.03 (0)	1.24 (0)	3.19 (0)	0.95 (0)	1.73 (0)
East-	light_var	4.40	3.81	0.98	10.21	6.49	0.06	1.13	12.10	0.50	4.41
Asian	acc_var	0.74	52.84	0.00	10.35	3.56	0.00	1.36	19.54	0.01	9.82
30-35	gyr_var	0.11	4.98	0.00	1.89	1.45	0.00	0.24	3.13	0.00	1.31
#9, female	MAE (drop%)	1.23 (0)	30.60 (0)	1.02 (0)	4.84 (0)	3.06 (0)	1.06 (0)	1.41 (0)	2.93 (0)	0.99 (0)	5.24 (0)
East-	light_var	0.30	0.93	0.23	0.37	0.18	0.06	0.72	3.24	0.04	0.68
Asian	acc_var	0.31	51.39	0.00	8.85	1.53	0.00	3.78	13.66	0.00	8.84
25-30	gyr_var	0.04	8.11	0.00	2.31	1.02	0.00	0.80	3.67	0.00	1.77
#10, male	MAE (drop%)	1.05 (0)	12.57 (47)	1.42 (0)	10.78 (0)	4.28 (2)	2.65 (0)	6.71 (0)	5.60 (0)	2.81 (0)	4.92 (5)
East-	light_var	0.25	3.36	0.41	3.70	2.56	0.56	2.12	2.53	0.18	1.74
Asian	acc_var	0.34	50.58	0.00	7.82	8.91	0.00	2.67	13.94	0.00	9.36
25-30	gyr_var	0.02	5.45	0.00	1.62	3.13	0.00	0.58	2.00	0.00	1.42
#11, female	MAE (drop%)	1.54 (0)	18.70 (0)	0.63 (0)	20.26 (0)	8.21 (0)	1.34 (0)	1.20 (0)	3.28 (0)	1.21 (0)	6.26 (0)
East-	light_var	15.69	115.21	1.80	34.98	19.81	0.16	3.37	26.57	4.12	24.63
Asian	acc_var	0.88	58.42	0.07	12.54	8.98	0.03	1.24	35.87	2.83	13.43
25-30	gyr_var	0.15	10.84	0.01	2.79	4.86	0.02	0.38	9.83	0.54	3.27
#12, male	MAE (drop%)	4.39 (0)	23.46 (0)	0.66 (0)	13.71 (0)	18.95 (0)	1.10 (0)	7.64 (0)	5.32 (0)	0.62 (0)	8.43 (0)
East-	light_var	2.70	15.18	0.36	4.19	3.64	0.03	0.95	4.06	0.08	3.47
Asian	acc_var	0.17	45.18	0.00	9.67	6.32	0.00	1.33	26.11	0.00	9.86
25-30	gyr_var	0.02	7.32	0.00	1.74	2.53	0.00	0.19	3.79	0.00	1.73
Overall	MAE (drop%)	4.55 (0)	22.60 (27)	1.17 (0)	12.13 (0)	9.28 (8)	3.10 (4)	3.71 (0)	5.03 (1)	1.71 (2)	
	light_var	6.88	384.06	2.37	26.85	11.44	0.15	6.69	111.70	0.58	
	acc_var	0.42	54.52	0.06	10.63	4.93	0.01	2.15	20.67	0.24	
	gyr_var	0.05	8.60	0.01	2.17	2.01	0.00	0.45	3.61	0.05	

Table 14: User study evaluation with Gear

participant-wise error based on the second group of main user study using Samsung Gear device.

From the last row of Table 14, overall error (MAE) ranges across activities, from 1.17 bpm (Sitting) with 0 drop rate to 23.39 bpm (Jumping) with 28% drop rate. This illustrates the HR monitoring performance is significantly influenced by the type of motion employed in different activities. Generally, intense activity and constant hand/arm motion result in significant HR measurement error, while lower error is observed during activities of intermediate/low level of motion. The highest error occurred during the *Jumping* activity, where the motion is the most intense and constant hand and arm motion takes place. Noticeable errors occur during *Folding* and *Walking* activities, 12.12 bpm with 0% drop rate and 10.06 bpm with 8% drop rate, respectively. During *Walking* activity, the hand and arm repeatedly perform a periodic movement of swinging forward and backward, while during *Folding* activity the users repeat a sequence of hand and arm movements. Though *Walking* activity comprises body motion while *Folding* activity does not, these two activities are similar as they both involve continuous hand and arm motion through the whole activity, and relatively high error is in presence in both activities. The compounded body motion does not impact the HR monitoring performance according to the comparison of MAE but it may be the cause of higher drop rate. The motion indexes *acc_var* and *gyr_var* of *Gaming* are the second highest among all the activities. However the MAE and drop rate are smaller than those of *Folding* and *Walking*. During the *Gaming* activity, a penalty shot was encountered at around 2 second intervals during the Kinect game. This, however, does not necessarily imply that the movement happened every 2 seconds because the penalty shots can be directed at the body and as a result the participant does not need to move. In addition, if the participant missed three shootings there was a short break of around 15 seconds before the next round started. These factors make the motion during *Gaming* discontinuous while the intermittent motion is still intense. Therefore, therefore the results suggests that the HR monitoring sensor seems to be more robust to discontinuous motion even the motion level is higher. During rest activities, *Standing*, *Sitting1* and *Sitting2*, the error is quite small and almost no drop down. Samsung Gear, as a newer device, provides better performance than the older devices, MSB and Fitbit Surge, especially for *Typing*, *Rubik* and *Gaming*. However, it still suffers from errors caused by motion in activities like *Jumping* and *Folding*. This suggests that motion and the noises caused by it still remains a challenge in PPG-based HR monitoring for the newer generation of wearable devices.

As shown in the last column of the Table 14, the overall error varies a lot across users as well. The MAE ranges from 1.9 bpm with 0% drop rate to 13.46 bpm with 19% drop rate, which indicates the individual diversity of participants also plays an important role on the HR accuracy in addition to the type of motion artifacts involved in different activities. This is mainly caused by 1) the inherent physiological characteristics of individuals, such as skin inhomogeneity [69], cardiovascular fitness, biological features [95], and 2) differences in the way the individuals perform the same activity, either in a trivial or non-trivial manner. Individual biological and physiological characteristics directly affect the quality of PPG signal that is the source of HR estimates on smart wearables. For example, the skin tone influence the penetration and absorption of the emitted sensor light, consequently influencing the reflected light that is used to derive the heart rate. Besides, participants perform the same activity differently in some cases. During two of the activities, *Typing* and *Rubik*, the hand motion intensity varies slightly across participants according to *acc_ var* and *gyr_ var*, while the MAE varies more significantly, ranging from 1.05 to 22.82 for *Typing* and from 0.81 to 9.44. This shows that the HR may differ noticeably across different individuals even when the motion levels are similar. During *Typing*, according to *acc_ var* and *gyr_ var*, the hand motion intensity varied only slightly among participants due to the different style and skill level of typing, while it was the same for *Rubik* activity during which the motion level remained more or less the same for different users. Moderate motion level can be found during *Folding* activity among participants as they generally followed the similar sequence of movements to fold the jacket but with trivial difference like whether they lift the jacket and/or how high they lift the jacket while folding. This is reflected in the variation of *acc_ var* (from 8.85 to 13.72) and *gyr_ var* (from 1.62 to 2.79) across participants as the HR errors also varies clearly. The motion pattern during *Jumping* activity also varies a lot across participants, which is mostly caused by different skill levels and style of rope jumping. For example, some participants tend to swing their arms with bigger magnitude while some others almost only move their wrist joint slightly without too much swing of arms. Experienced participants can keep the jumping more consistent with less breaks, resulting in higher motion level compared to those who took more breaks as the rope hit their feet at times. For *Jumping*, *acc_ var* ranges from 31.14 to 119.45, which also results large differences in HR errors across participants. The variation observed in motion levels and HR errors across participants for activities like *Walking* and *Jumping*, suggests that motion patterns and HR errors could be very different even for a same activity. Based on the participant-wise analysis,

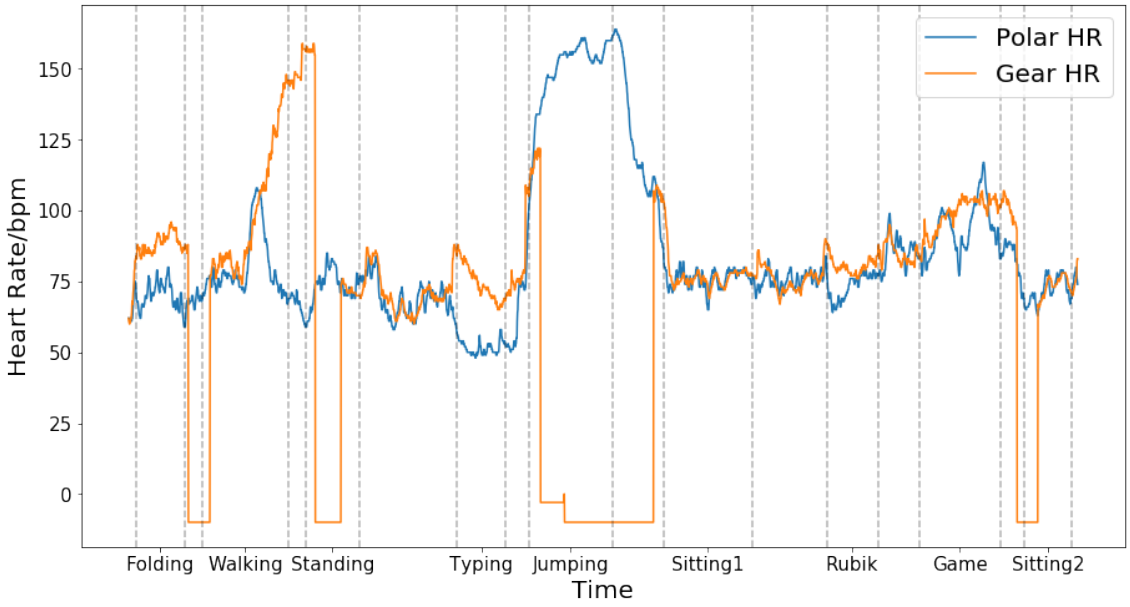


Figure 9: Example: detailed measurements of participant 3

we can find that motion levels and HR errors can vary considerably even for the same activity due to the user diversity. This shows the difficulty of developing a generative HR correction model for a variety of users. In Section 6, we introduce our deep learning based approach to tackle this challenge.

5.5 Analysis of HR Monitoring Failure

We illustrate how HR monitoring on wrist-worn wearables may fail in everyday usage through an example in our main user study, shown in Figure 9. Motion can cause the HR monitor to lose track of the correct HR and the monitoring keeps failing even after the activity has stopped. For example, the device gives erroneous HR measurement for a while after the *Walking* activity where HR monitor initially starts to fail, followed by a drop down period before it tracks the correct HR again. Although there is a break period before the *Standing* activity, the Gear watch fails mostly during the *Standing* activity, causing significantly higher error and drop rate compared with other participants. Also the similar phenomenon happens for the participant during *Sitting2* activity where the monitoring failure from the previous activity causes a moderate drop rate of 29%. Actually, it also takes a while after the *Jumping* activity for the Gear to follow the correct HR. However, it does not influence the following *Sitting1* activity as longer break is given after the intense *Jumping* activity. These failures happens more often for this

participant most likely because of his special individual physiological characteristics. Nevertheless, the failures observed from this case illustrate vulnerability of PPG-based HR monitoring in everyday usage.

5.6 Summary

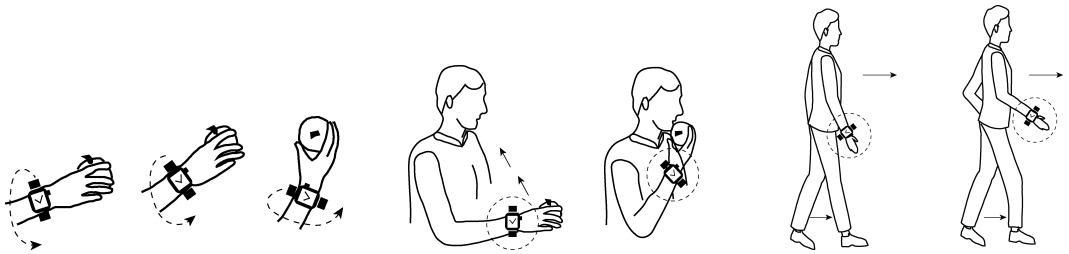
The results of our user studies were illustrated and analyzed thoroughly in this section. There results are summarized as follows.

- The result suggests that the heart rate monitoring by the PPG-based wearable lacks accuracy under the everyday scenario.
- We build the motion index (acc_var and gyr_var) to quantify the motion, followed by further analysis of the HR measurement errors with respect to the motion index. Intermediate to high intensity motions cause significant errors in HR monitoring, while the activities of low physical intensity but trivial and irregular hand/wrist motions usually result in noticeable errors.
- Tight strap attachment to the wrist is helpful to mitigate the consequence of motion in HR monitoring, while looser strap usually induce significant error in the HR measurements.
- We discuss the activity-wise and participant-wise HR measurement errors to show the HR monitoring performance of the wearables varies significantly across different participants.
- We illustrate the vulnerability of HR monitoring on wrist-worn devices in everyday use with an detailed example in our user study.

According to our analysis in this section, current wearables fails to offer accurate heart rate monitoring, with motion being a major source of the errors. In next section, we propose our solution DeepHR for calibrating HR errors.

6 DeepHR: Deep Learning Based Heart Rate Calibration

We have shown that the HR monitoring on wrist-worn wearables is prone to motion and errors caused by it. This reduces the applicability of these devices, especially



(a) Wrist motion (b) Hand + wrist motion (c) Body motion

Figure 10: Motion patterns that affect the accuracy of heart rate measurements. The patterns are transformed into sensor representations as part of the DeepHR sensing pipeline.

when used pervasively in everyday life. To address this issue, we propose DeepHR as a novel technique to calibrate PPG-based heart rate monitoring on wrist-type wearables. High quality and accuracy of the continuous heart rate monitoring are crucial. However, the heart rate monitoring provided by off-the-shelf commodity devices fails to offer adequate performance as shown by our assessment in Section 5. By improving the accuracy of the heart rate measurements from the commodity devices, DeepHR supports and enables those researches relying on wrist-type wearables for continuous heart rate monitoring in everyday use. The key idea of DeepHR is to study the relationship between HR errors and motion, taking the temporal dependency of the time-series data into account. The HR error is obtained by calculating the difference between PPG-based HR measurements and a reference device. The captured motion measurements and PPG-based heart rate measurements are given as input to a deep learning model that learns a calibration function, which in turn can be used to predict the difference between PPG heart rate and the reference heart rate. As demonstrated in Section 5, the relationship between motion and heart rate measurement errors are complex and cannot be captured by naive solutions. By employing deep learning, DeepHR is capable of learning this complex relationship. As we demonstrate in Section 7, DeepHR can improve accuracy significantly. In this section, the DeepHR are described in detail, including the underlying motion capture mechanism, and the structure of the deep learning model.

6.1 Motion Capture on Smart Wearables

In this section, we introduce how the motion information is quantified and describe how the motion representation is extracted from the sensor measurements.

Motion Patterns In Figure 10, we illustrate how 3 types of typical human motions that generally influence the HR monitoring accuracy are characterized on wrist-worn sensors. The idea in DeepHR is to simplify the motion complexity by focusing on particular type of motion. Figure 10a shows how the wrist motion is represented on the devices, which is mostly captured by gyroscope, while Figure 10b shows the combined wrist and hand motion that is mostly characterized by accelerometer and gyroscope, and Figure 10c illustrates body motion that affects both gyroscope and accelerometer. Note that sustained motion that is not caused by the user, for example riding on a bus, can also influence both gyroscope and accelerometer. While DeepHR does not provide a separate component to differentiate this type of motion from the 3 types of motion discussed before, the representations abstracted from accelerometer and gyroscope are capable of identifying sustained motion patterns according to previous studies [33].

Measurements and Preprocessing DeepHR utilizes measurements from the tri-axial inertial sensors (accelerometer and gyroscope), together with two types of simultaneously collected heart rate measurements, collected from PPG device that needs to be calibrated and the reference ECG device, respectively. The inertial sensors are used to capture motion, and these are associated with HR error between the PPG device and the reference. We use MSB and Samsung Gear as the PPG devices and Polar H7 as the reference sensor in our study. The HR sampling rate of Polar H7 and Samsung Gear is 1Hz while the HR sampling rate of MSB is varying, usually between 1 second and 1.5 seconds. The motion measurements from accelerometer and gyroscope are collected from MSB with a frequency of 62.5Hz, from Samsung Gear with a frequency of 100Hz. Before the data collection started, all the devices were synchronized according the network time protocol (NTP) so that measurements from different devices could be matched and aligned. After the data collection, the measurements were interpolated to have a consistent sampling frequency using linear interpolation, with 1Hz for the two types of heart measurements and 50Hz for the tri-axial inertial sensors, and all the measurements are aligned according to the synchronized timestamps. We consider the data collected under two different scenarios evaluate DeepHR. Firstly, we use data collected from the main user study in Section 4.1. Secondly, we use the data collected in prolonged everyday scenarios (discussed in Section 7.1). For the first set of collected data, we split it into segments of activities with the transition period and any invalid measurements (e.g., drops) removed.

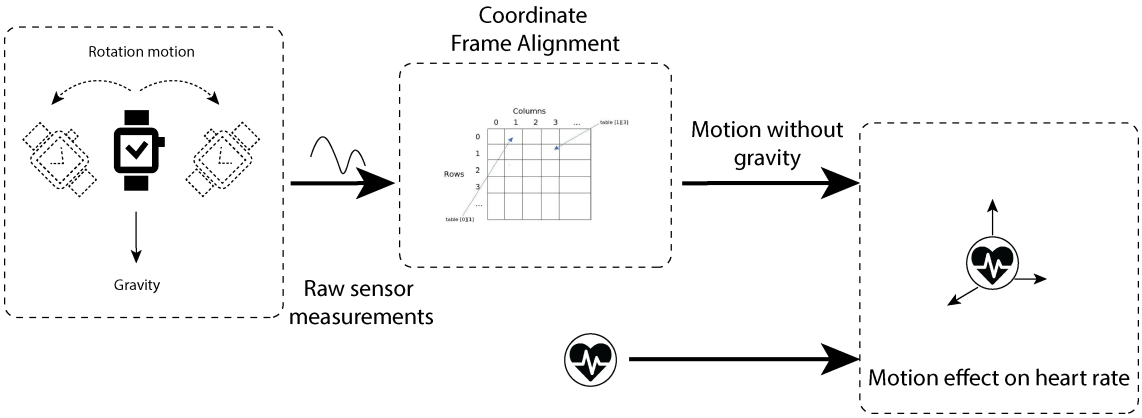


Figure 11: Quantification of effect of motion in heart rate.

Coordinate Alignment The motion measurements collected from the tri-axial motion sensors on the wearable are affected by orientation of the device and the gravity. These two factors combined can sometimes makes it difficult to extract the true motion information if no effective processing is performed on the raw sensor measurements. To capture the motion information, we first align the raw measurements with a global coordinate system [34, 66]. A non-linear complementary filter [61] is applied to estimate the gravity component utilizing measurements from accelerometer and gyroscope. This technique was chosen because it balances the trade-off between accuracy, simplicity of implementation, and runtime performance [66]. Then the coordinate system is aligned with a global coordinate system based on the estimated gravity component. We eliminate the gravity component to obtain the linear acceleration after the coordinate is aligned. The linear acceleration is not influenced by gravity and better describes the motion information.

Representations After the coordinate alignment, the gravity-free linear acceleration consists of two non-directional horizontal components and a directional vertical component. Horizontal linear acceleration captures the lateral and longitudinal movement of the device, which can be caused by both the body and hand motion. The vertical linear acceleration component describes the upward and downward movement of the device. Although neither the horizontal linear acceleration nor the vertical linear acceleration can distinguish the hand and body movement, the relationship between the two can shed light on the motion pattern and provide clues on which one is dominant. For example, the horizontal and vertical accelerations are highly correlated during swinging motion, while one component illustrates the dominance over another during the body movement (without heavily compounded

with hand motion). We also rotate the gyroscope measurements to achieve a motion representation, which provides information about the rotational movement of the the device. In DeepHR, the deep learning model takes the linear acceleration and rotated gyroscope measurements as input to extract the motion information that relates to HR error.

Frame Formulation A frame consisting of motion and heart rate data for the last 10 seconds, including the linear acceleration, rotated gyroscope measurements, and heart rate, is formed to predict the current HR error by the wearable. This enables the calibration model to take temporal dependencies in the measurements into account. As the motion is known to be tightly related with the HR value and its variation, HR typically rises with some delay instead of increasing instantaneously with a jump. Also the duration of the motion is associated with the severity of its corruption on the PPG-based HR estimates, the longer the duration the more likely the HR estimates become erroneous. Beside, on current wrist-type PPG wearables, motion correction techniques are integrated as well as some filters, to calculate the heart rate. These internal compensation mechanisms of the wearables also induce delays to the effect of motion on HR value and its error. For example, temporal smoothing like moving average with a sliding window is often applied to calculate the heart rate, resulting in a delay between the happening of the motion and observing the corresponding effect. Therefore, forming a frame that includes the necessary temporal motion and heart rate information is crucial for accurate predictions on HR errors. Considering the computational power needed to process the frame, and the trade-off between incorporating more temporal information in the frame and enabling quick response to changes, in our implementation 10 seconds is chosen as the frame length.

6.2 Learning Calibration Function

DeepHR learns a calibration function through a deep learning model. The model takes takes motion data (i.e., linear acceleration and rotated gyroscope) described in Section 6.1 and the heart rate measurements from the PPG wearable device as input, and outputs an estimate of the HR error. During the learning phase, the model needs to have reference heart rate measurements, for example Polar H7 used in our study. The estimate of the HR error is then used to compensate the heart rate measurements from the wearable to obtain a calibrated heart rate measure-

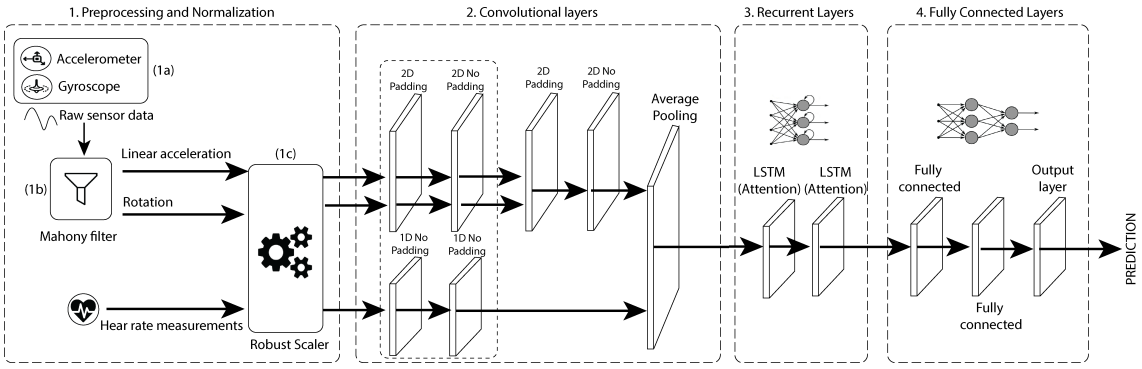


Figure 12: The deep neural network structure in DeepHR.

ments. The deep learning model used in DeepHR comprises of 10 layers, including 4 convolutional, 2 recurrent, 2 MLP, and 2 pooling layers. The overall structure of DeepHR is shown in Figure 12, and the details of the model is introduced in this section.

Preprocessing and Normalization Preprocessing and normalization is performed on the data before it is passed as input to the model. In Section 6.1, the necessary preprocessing steps have been described, including the interpolation to solve the inconsistent sampling rate of sensors, synchronization to match the different frequencies of different sensors, coordinate alignment, building motion representations, and frame formulation. Besides these steps, a normalization step is needed to mitigate the influence of outliers and to standardize the range of different input dimensions by scaling. In DeepHR, we normalize measurements using a robust scaler is applied to the input data by removing the median and scale the data using the interquatile (IQR) on each dimension of the input data. The scaling parameters (the median and IQR) are learnt during the training phase and store for future input to the model.

Convolutional Layers Convolutional layers are a specialized deep learning structure for grid-like data and we use convolutional layers in our study to process time-series data. In DeepHR, the convolutional layers take the preprocessed linear acceleration, rotated gyroscope measurements, and heart rate measurements from the PPG-based wearables as input, to extract effective feature representations that are later associated with the HR error. The 3-dimensional linear acceleration and gyration input first goes separately to two consecutive $2D$ convolutional layers which fuse and extract features capturing intra-sensor characteristics. The filter size of

the first layer is 25×3 and the convolution is done without padding, resulting in the same output shape as the input, which extracts the intra-sensor features from the input and keeps the dimensions of the input. Then the second convolutional layer operates with a filter of the size 25×3 without padding, which further extracts intra-sensor features and reduces the 3-dimensional input into a single dimension because no padding is applied. The first dimension of the filter is designed to cover the measurements of half a second, which corresponds to 25 for MSB and 50 for Samsung Gear, while the second dimension is designed to be 3 corresponding to the three dimensional linear acceleration and rotated gyration. Note that the filter size is changed to 50×3 accordingly for Gear collected data as the frequency is 100Hz. Once linear acceleration and gyration have passed through the first two convolutional layers, another two convolutional layers with filter size of 25×2 (with and without padding) are applied to the concatenated output of linear acceleration and gyration to repeat the similar process. These two convolutional layers fuse and extract the inter-sensor features, then reduce the output into single dimension. The second dimension of the filter corresponds to the two dimensional concatenated linear acceleration and gyration features output from the first two layers.

In summary, we first extract 3-dimensional features from both linear acceleration and gyration in the first and second convolutional layer. Next, the extracted features are mapped into single dimensional features through the third convolutional layer and concatenated to form a two dimensional input for following layers. Next, we extract two dimensional features through the third convolutional layer. Then the fourth convolutional layer to perform the dimensionality reduction, resulting in a single dimensional feature combining all the motion information. In parallel, the heart rate measurements pass through two 1D convolutional layers with a filter size of 25×1 (no padding). Average pooling is applied to the HR feature and unidimensional motion feature resulting from the corresponding convolutional layers, to make the model more robust to noise and distortion in data [54]. The pooled motion information and HR information are then concatenated. All the convolutional layers use rectified linear unit (ReLU) as activation function and the number of filters is 32, i.e., 32 different feature layers are generated from the input.

Recurrent Layers The concatenated output from convolutional layers is connected to the recurrent layers, specifically the long short-term memory (LSTM), to capture temporal dependencies in the measurements across multiple frames. Two LSTM layers with attention mechanism [39] are stacked together by returning the

intermediate recurrent outputs from the first LSTM layer as the input for the second LSTM layer. The attention mechanism calculates a weight vector that evaluates the significance of intermediate output at each individual time step. The final output is given as a dot product of the weight vector and the LSTM intermediate output vector produces the final output, which allows finer tuning and combination of each time step. 10% of the intermediate recurrent state is dropped out randomly to increase the robustness and the generalizability of the model. To achieve a balance between the capability of the recurrent layers and the computational overhead, the number of units for the LSTM layer is set to 128 for both layers.

Fully Connected Layers Finally, 3 fully connected layers (MLP) take the output from the recurrent layers and further produce the output as the HR error estimate for compensation. The first two MLP layers have 64 and 32 units, respectively, and dropout of 10% is applied after the first MLP layer. The last layer is the output layer, which has only one unit, resulting in a single value as the estimated HR error as the final prediction.

6.3 Summary

In this section, we introduced the design of the DeepHR. First, we illustrated that motion can be roughly divided into 3 categories (hand, wrist, and body motion) and combinations of these, followed by description of how to quantify and extract these motion patterns from the accelerometer and gyroscope on the wearables. Then, we discussed the deep learning model that takes frames of the motion representation and PPG heart rate as input to output the estimated HR error. The deep learning model consists of preprocessing and normalization, convolutional layers, recurrent layers, and fully connected layers, in total 4 components as shown in Figure 12. In preprocessing and normalization, the input data is scaled by a robust scaler to mitigate the effect of outliers and to make the input data suitable for the deep learning model by converting the different source of input data values to a similar range of scale. The convolutional layers effectively extract features from the frames of motion representations, while the recurrent layers model temporal dependencies in the time-series data, and finally the fully connected layers output a single value as the prediction for the HR measurement error. We next discuss the performance of DeepHR using data collected from the user study described in Section 4.1 and data collected in uncontrolled everyday using scenario. The DeepHR performance is

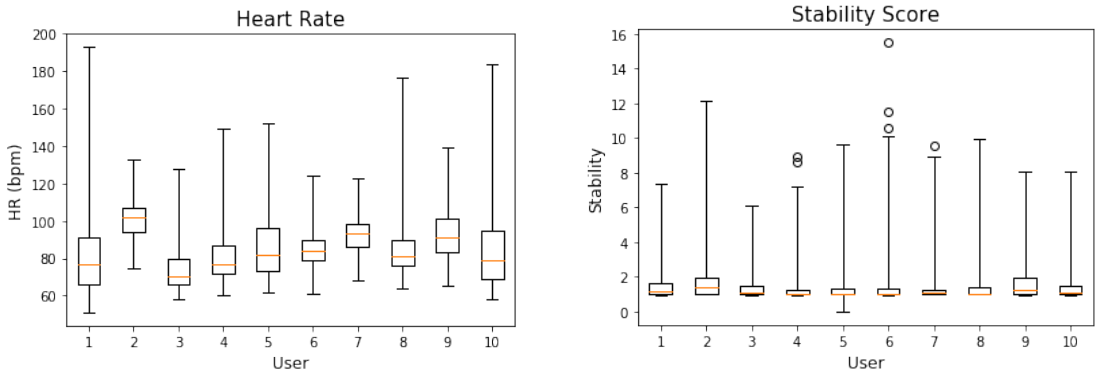
benchmarked against other techniques to show the good performance of DeepHR.

7 DeepHR Performance

We validate and evaluate DeepHR with the user study data and uncontrolled everyday data that are introduced later in Section . DeepHR is benchmarked against other approaches based on conventional feature engineering technique. In this section, we first elaborate on the evaluation scenarios and procedure, and then discuss the DeepHR performance in both everyday use and user study scenarios. Finally, we compare the performance of DeepHR with other benchmark approaches.

7.1 Evaluation Scenarios and Procedure

As our main source of evaluation data, we collected heart rate and motion measurements in everyday use of the wrist-worn heart rate monitoring devices from 10 participants (average 25.8 years old, SD=3.4, 5 females). The participants for everyday data collection are different from the participants employed in our user study described in Section 4.1. During the data collection, participants wore the Microsoft Band 2 (MSB) on their dominant hand to collect the PPG-based HR, and a Polar H7 strap on the chest to collect reference data simultaneously. The devices being used are illustrated in Figure 7a, and the device setup is shown in Figure 7b. In addition, a Samsung S6 smartphone was carried by the participant to store the data that was measured by MSB. Before the data collection started, the participant was instructed on how to wear the devices properly and on how to initialize the data logging correctly. Participants were asked to wear the devices and record data whenever it was comfortable and convenient for them. The participant was asked to collect data from regular daily activities, but without any specific instructions on preferred types of activities. Due to the limitation of battery, maximum period for continuous collection is around six hours. However, most participants only collected continuous data for around $3-4$ hours without interruption because they consider wearing the heart rate chest strap for a prolonged period uncomfortable. Each continuous set of measurements is considered as a segment. During training and testing, segments from a same participant are never mixed into both training and testing data sets at the same time to ensure the validity of our validation setup [30].



(a) Distribution of measurements across participants in the everyday data.

(b) Distribution of measurements across participants in the everyday data.

Figure 13: Summary of heart rate and motion measurements across participants in the everyday dataset.

Data Cleansing Due to sensor and logging failures, data cleansing need to be applied to the collected data before it is utilized. Logging failures can happen when the sensors lose connection with the smartphone, e.g., it is away from the participant. Poor connectivity between the sensor and the smartphone due to extended distance can cause errors in the recording. In addition, sensor errors are highly likely to occur during prolonged data collection process, considering the discomfort (mostly the chest strap) and diversity of activities, which can cause unexpected sensor displacement and bad contact between the sensors and the measurement sites. In our data collection, we only consider data from periods when all measurements (i.e., HR and motion) are available and valid. As a result, the total duration of available data from each participants varies significantly, ranging from 1.67 to 16.62 hours.

Data Diversity In Figure 13, the distribution of heart rate measurements from the reference Polar H7 and motions across different participant are shown to demonstrate the diversity of heart rate measurements and motions. From Figure 13a, the differences of heart rate measurements distribution between different users are notable. To assess the diversity of motions, we use stability score [34] that combines acceleration and gyration to indicate the level of motion. Higher stability score refers to higher level of the activity in general. According to the distribution of motion measurements shown in Figure 13, participants indeed performed diverse daily activities during the data collection, and the activity pattern and intensity

Time Domain	mean, median, standard deviation, variance, range, max, min, RMS, zero crossing, IQR, integral, double integral, cross correlation, auto correlation, difference
Frequency Domain	FFT DC, FFT spectral energy, FFT peak frequency, 0Hz energy, 0-0.8Hz energy, 0.8-2.5Hz energy, 2.5-10Hz energy, 10Hz+ energy
Representations	Linear acceleration (3-axes), rotated gyration (3-axes), pitch angle, roll angle, magnitude of rotated acceleration, magnitude of rotated horizontal acceleration, magnitude of rotated vertical acceleration, magnitude of linear acceleration, magnitude of rotated gyration.

Table 15: List of features and representations considered by baseline techniques.

vary across different participants.

Comparisons To demonstrate the effectiveness of the convolutional and recurrent modules, we benchmark DeepHR against two other models that rely on conventional feature engineering. The first baseline model is a conventional feedforward neural network that only consists of MLP, and the second is a deep learning model that combines LSTM and MLP. By comparing DeepHR to the two baseline models, we demonstrate the effectiveness of convolutional module and recurrent module in DeepHR.

Feature Engineering: The features being input to the baseline model were designed according to the state-of-art motion sensing techniques [21, 33], and all the features we considered in the baseline model are described in Table 15. First, 13 different representations (Table 15) are extracted from the raw accelerometer and gyroscope measurement data. Next, time and frequency domain features are extracted from the representations to capture a wide range of possible motion patterns. Similar to DeepHR, the baseline models utilize linear acceleration and rotated gyration. Besides these, the baseline models consider several additional representations: the pitch and roll angles of the wearable device derived from Mahony filter (described in Section 6.1), and the magnitude ($L2$-norm) of the rotated acceleration, horizontal linear acceleration (i.e., $L2$-norm of the x and y axis of the rotated measurements), rotated vertical acceleration (i.e., z axis of rotated linear acceleration), magnitude of

gravity eliminated linear acceleration, and magnitude of the rotated gyration. For the first baseline model with only MLP, the input features were extracted using a once second window with 50% overlapping in the adjacent time-series measurements as this resulted in the best performance. For the second baseline model, the time-series measurements were first segmented into frames of 10 seconds to match the input shape required by recurrent module, enabling the second baseline model to take temporal dependency into account in contrast to the first baseline. In total, 301 features were derived to be used as input with the two baseline models. The input features together with the HR were input to the baseline models to predict the estimate of HR error between the wrist-worn HR monitor and the reference device. In the following two paragraphs, we introduce the detailed structures of our two baseline models.

Deep NN: The first baseline model consists of a deep feedforward neural network. Similar to DeepHR in term of depth, it contains 9 hidden layers (same as that of DeepHR). However, the total number of layers is smaller than in DeepHR because pooling layers are not necessary as the motion features are handcrafted instead of extracted through convolution. The first hidden layer and the output layer use linear function $f(x) = x$ as activation, whereas the other layers use either hyperbolic tangent (tanh) or rectifier linear unit (RELU) activation. Specifically, with tanh for the first, third, and fourth layer. After the third, sixth, and ninth layers, a dropout of 0.1 is applied to prevent overfitting. The number of neurons in the layers are 301, 512, 1024, 512, 1024, 512, 256, 128, and 64 respectively.

Deep NN + LSTM: Our second baseline model consists of feedforward and recurrent neural networks, similar to the first baseline model but taking temporal dependencies into account. Overall, the network consists of 4 fully connected hidden layers that are wrapped by time distributed wrappers, two LSTM layers, another two layers of feedforward network, and an output layer, with 0.1 of dropout performed after the third and eighth layer. The structure of the second baseline model is similar to the first baseline model except that the fifth and sixth hidden layers are substituted by two stacked LSTM layers (0.1 recurrent dropout applied) of 256 and 128 units respectively, and all the layers preceding the LSTM layers are wrapped by a time distributed wrapper to be compatible with the recurrent feature of LSTM. To incorporate temporal dependencies, the input features are segmented into frames of 10 seconds. The input share is $(20, 301)$, where 20 corresponds to a 10-second frame and 301 corresponds to the dimension of hand-crafted features. With feature

		Deep NN		Deep NN+LSTM		DeepHR		Duration(h)	
User	**Original**	**Prediction**	**% Gains**	**Prediction**	**% Gains**	**Prediction**	**% Gains**	**Training**	**Validation**
1	10.54 (SD=13.05)	14.06 (SD=9.23)	-33.40 (29.28)	7.65 (8.88)	27.38 (32.00)	7.60 (SD=8.95)	27.87 (31.40)	70.33	10.72
2	20.61 (SD=13.43)	10.99 (SD=8.18)	46.67 (39.08)	12.16 (9.60)	40.99(28.52)	9.99 (SD=8.14)	51.51 (39.40)	76.76	4.30
3	6.31 (SD=8.48)	15.32 (SD=6.41)	-142.94 (24.39)	6.41 (7.07)	-1.59 (16.58)	7.63 (SD=7.04)	-21.08 (16.95)	79.38	1.67
4	6.95 (SD=9.56)	5.70 (SD=6.65)	18.04 (30.44)	6.35 (7.06)	8.62 (20.52)	5.42 (SD=6.13)	21.99 (35.89)	70.85	10.20
5	9.06 (SD=11.57)	16.01 (SD=6.90)	-76.70 (36.46)	6.14 (6.96)	32.27 (39.82)	5.63 (SD=6.79)	37.84 (41.32)	68.35	12.71
6	7.22 (SD=6.86)	16.71 (SD=6.90)	-131.48 (-0.48)	4.84 (4.57)	33.03 (33.40)	4.84 (SD=4.77)	32.98 (30.55)	64.44	16.62
7	13.10 (SD=9.91)	12.27 (SD=7.86)	6.33 (20.70)	10.48 (8.61)	19.88 (13.13)	7.92 (SD=7.07)	39.55 (28.72)	73.66	7.40
8	9.17 (SD=13.18)	11.78 (SD=10.06)	-28.45 (23.66)	5.37 (6.76)	41.47 (48.70)	6.64 (SD=8.26)	27.59 (37.32)	76.75	4.30
9	15.09 (SD=13.46)	12.02 (SD=7.77)	20.32 (42.24)	8.39 (8.96)	44.41 (33.39)	7.02 (SD=7.83)	53.51 (41.83)	77.16	3.89
10	9.66 (SD=14.55)	15.20 (SD=8.46)	-57.31 (41.87)	6.96 (9.82)	27.99 (32.51)	6.98 (SD=10.46)	27.80 (28.09)	71.80	9.25
Overall	10.77 (SD=11.40)	13.01 (SD=7.89)	-37.89 (28.76)	7.47 (7.88)	27.45 (29.86)	6.97 (SD=7.54)	29.96 (33.15)	72.95	8.10

Table 16: Performance evaluation of DeepHR and baseline models with everyday data.

engineering, we use a one-second sliding window with 50% overlapping, 20 frames of the extracted features covers a period of 10 seconds and each frame is a single timestep in LSTM. Two feedforward layers (with tanh and RELU as activiation function, respectively) follow the two LSTM layers, after which the final output is calculated.

7.2 Performance in Everyday Scenario

We first demonstrate that DeepHR can notably mitigate the heart rate measurements error, and that it can generalize across different users and activities. This is achieved through a leave-one-user-out cross validation using the everyday data described in Section 7.1. The model is trained with 9 users and validated with the remaining user, with each participant used for validation once, resulting in 10 folds cross validation. The average across 10 participants is reported as the overall performance.

The results of the evaluation are described in Table 16. The column *original* shows the difference between the PPG-based device (MSB) and the reference sensor (Polar H7), and the prediction columns shows the difference (mean absolute error) between reference measurements and the predictions of the different models (Deep NN, Deep NN + LSTMj, and DeepHR). The column *%Gain* shows the gain in percentage. The last two columns illustrate the duration of data used for training and validation, respectively. As shown in Table 16, DeepHR improves the accuracy for all the participants except participant 3 who has the least amount of available data shown in the column *validation*. The overall improvement by DeepHR is close to 30 percent, reducing the error in HR measurements from 10.77 bpm to 6.97 bpm. DeepHR

also decreases variation in the heart rate errors, as can be seen from the decrease in standard deviation. The original HR error across users varies considerably, ranging from 6.31 bpm to 20.61 bpm, while the HR error only ranges from 4.84 bpm to 9.99 bpm. Also, the standard deviation decreases by more than 30 percent. As the training and validation data was collected from different users and the daily activities performed by them are likely to be diverse, DeepHR performs well across the different validation users, which demonstrates its capability of generalizing in everyday scenario.

DeepHR has been benchmarked against other two baseline models and the result is described in Table 16. The result demonstrate the advantage of DeepHR over the conventional feature engineering techniques and its capability due to the convolutional and recurrent modules. In the everyday use scenario, the performance of DeepHR is better for most of the user (7 out of 10) when compared to Deep NN + LSTM, and better for all the users when compared to Deep NN. As both the Deep NN + LSTM and DeepHR consider the temporal dependencies, the improvement of the performance of DeepHR comes mainly from the convolutional module that effectively extracts features from the sensor representations instead of using conventional feature engineering techniques. Additionally, the poor performance of Deep NN suggests it is necessary to incorporate temporal dependencies.

7.3 Performance in User Study

Next, we validate DeepHR with data from 12 participants collected from our user study described in Section 4.1 to understand how the model performs on different types of activities representing motions present in everyday activities. Mean absolute error (MAE) is used as evaluation metric, and is first calculated for an activity for a single user, resulting 12 MAE values corresponding to the 12 participants. These are then averaged into a single MAE to represent the error of a particular activity. To test the generality and how similarity of training data affects performances, three different sets of training data were utilized to train DeepHR and baseline models while the validation data is always the user study data. These sets are: user study data, everyday data, and the combination of everyday and user study data. The first evaluates the model performance when the target activities are the same as in the training data. The second assesses the generalizability of DeepHR to activities that are not necessarily present in the training data. Finally, the third one explores the how the model can adjust itself when a small amount of data from target activities

		Deep NN			Deep NN + LSTM			DeepHR		
Activity	**Original**	**Controlled**	**Everyday**	**Combined**	**Controlled**	**Everyday**	**Combined**	**Controlled**	**Everyday**	**Combined**
Typing	7.17 (2.86)	7.45 (3.92)	7.20 (4.00)	10.47 (4.76)	7.83 (3.71)	45.85 (3.59)	6.89 (3.23)	7.37 (4.05)	7.80 (4.30)	8.09 (4.07)
Rope Jumping	68.50 (8.69)	25.90 (13.95)	54.54 (9.70)	46.52 (9.61)	22.74 (11.48)	21.45 (7.49)	31.26 (12.24)	20.97 (11.75)	50.54 (8.64)	42.26 (10.57)
Lying Down	2.75 (1.45)	3.02 (2.34)	3.31 (1.88)	6.64 (3.29)	4.45 (3.18)	47.02 (2.88)	2.62 (1.53)	3.61 (1.92)	2.72 (1.78)	2.66 (1.55)
Folding Clothes	15.65 (4.55)	13.84 (7.71)	8.43 (5.22)	10.31 (5.32)	12.09 (6.93)	37.82 (5.06)	11.04 (6.20)	11.84 (6.68)	8.23 (4.79)	10.09 (4.89)
Indoor Walking	26.61 (11.45)	15.33 (9.19)	16.66 (9.52)	14.86 (9.43)	15.83 (9.84)	27.14 (10.56)	16.24 (9.78)	13.59 (8.90)	14.57 (8.57)	13.63 (8.25)
Standing Still	3.68 (3.92)	4.41 (4.89)	3.99 (3.65)	6.22 (4.30)	9.66 (5.19)	43.01 (5.18)	3.53 (3.78)	3.78 (3.89)	4.47 (3.97)	4.10 (3.65)
Rubik Cube	11.92 (2.93)	12.67 (4.99)	9.24 (4.49)	10.26 (4.39)	11.13 (5.14)	41.76 (3.06)	11.25 (3.73)	10.30 (4.80)	8.97 (3.25)	11.22 (3.63)
Motion Game	32.05 (9.67)	18.68 (11.61)	22.20 (9.23)	19.18 (9.74)	18.74 (11.81)	24.10 (8.71)	18.80 (11.23)	16.16 (10.38)	20.36 (8.66)	18.74 (8.82)
Sitting on a Sofa	2.03 (1.80)	2.16 (2.34)	2.44 (1.88)	4.83 (3.19)	2.85 (2.69)	46.21 (3.40)	1.86 (1.70)	1.92 (1.78)	1.96 (1.79)	2.07 (1.71)
Overall	18.93 (5.26)	11.50 (6.77)	14.22 (5.51)	14.36 (5.43)	11.70 (6.68)	37.15 (5.55)	11.50 (5.93)	9.95 (6.02)	13.29 (5.09)	12.54 (5.24)

Table 17: Performance evaluation of DeepHR and baseline systems with user study study. As training data we consider three variants: user study only (controlled), everyday data only (everyday), and combined everyday and user study data (combined).

is present in the training data. In the first setup, the models are trained with all the everyday data, and validated by considering all user study data as test data and overall error is macro-averaged across the activity-wise errors. In the second setup, the model is evaluated using the leave-one-user-out cross validation, i.e., the model are trained with the user study data from 11 users and validated with the remaining user until every user has been used for validation once. The third setup is otherwise analogous except that the data from everyday data is included in the training data. Note that the participants employed in the user study and in the everyday data collection are different from each other, and our evaluation scenarios ensure the data from a same user is never utilized in both training and validation data.

The evaluation results are illustrated in Table 17 and demonstrate superior performance of DeepHR compared to the baseline models. In Table 17, the column *original* shows the heart rate measurement error for each activity in the user study. The columns *controlled* (first setup), *everyday* (second setup), and Combined (the third setup) show the performance of different models under the aforementioned three different evaluation setups. The performance of DeepHR is the best when the training data consists of only user study data, with the error decreasing from 18.93 bpm to 9.95 bpm, and DeepHR improving the accuracy of heart rate measurements for most of the activities (7 out of 9). This demonstrates that DeepHR can capture the relationship between motion and heart rate measurement errors, and improve the accuracy of the HR monitoring. When only everyday data is used for the training (second setup), the performance gain is the smallest with an improvement of around 30% (overall error decreasing from 18.93 to 13.29). Compared

	MAE		Motion		Duration (h)		
User	**original**	**DeepHR**	**acc_var**	**gyr_var**	**Train**	**Validation**	**hr_var**
1	4.64	4.57	2.47	0.64	36.7	9.8	198
2	1.64	2.32	1.31	0.36	37.4	9.0	86
3	2.51	7.18	1.18	0.25	37.8	8.7	64
4	3.43	3.46	2.67	0.49	38.4	8.1	78
5	3.35	3.65	0.73	0.16	36.7	9.8	74
Overall	3.11	4.24	1.67	0.38	37.4	9.1	100

Table 18: DeepHR on everyday data collected from Samsung Gear S3.

to the first evaluation scenario, the performance of the model decreases clearly for activities of intensive motion, such as rope jumping and motion game, which were unlikely to be present in the everyday data. This seems to suggest that DeepHR struggles for activities that have significantly different motion characteristics than what is included in the training data. However, for activities that resemble common wrist and hand motion in daily life, like folding clothes and playing Rubik's cube, DeepHR can still provide improvement. In the third setup, the performance is slightly improved compared to the second evaluation setup as the user study data has been added to the training data. This gain is mostly from the activities of intense motion and suggests the model can be fine-tuned by adding the data from specific activities. Even if only a small amount of additional dat is included, it can be helpful to improve the performance. Indeed, the total duration of data in the user study data is only around 5.8 hours compared to 81.05 hours in everyday data. Though Deep NN + LSTM provides better performance than DeepHR in the third setup, the improvement is mainly from the most intense activity rope jumping. In addition, Deep NN + LSTM gives very poor performance while training in the second setup, which suggests overfitting and poor generalizability. Overall, DeepHR offers the most robust performance across all the evaluation scenarios.

7.4 Performance on an Additional Dataset

An additional data set was collected in everyday using scenario from 5 participants (3 males and 2 females) using Samsung Gear S3, to further test DeepHR performance. The everyday data is collected using the same scheme as described in Section 7.1, with the exception that Samsung Gear S3 is substituted for MSB. As shown in Table 18, DeepHR fails to improve the accuracy of the heart rate measurement.

There are two main reasons for this. The first is that HR accuracy of Gear (overall error 3.11) is much better than MSB (overall error 10.77 bpm shown in Table 16) in everyday scenario. Once the original measurement error gets smaller, it is more difficult to identify the situations where heart rate measurements are off because the error is much more trivial. The column *motion* highlights the intensity of motion in the data using motion index described in Section 5.2), which represents the overall motion intensity of the user's daily activity level. The mean motion indexes of the 10 users in the everyday data collected with MSB are 2.54 (acc_var) and 0.66 (gyr_var) while the value for the everyday data collected with Gear are 1.67 (acc_var) and 0.38 (gyr_var). This suggests the daily activity level is lower during the data collection with Gear, which can also be observed from the lower heart rate variation of the user during the data collection with Gear. We use hr_var, the variance of the reference heart rate measurements (Polar H7), to show the overall variation of the heart rate variation user across the whole data collection period. The mean hr_var during the everyday data collection with MSB is 270, while that during data collection with Gear is 100. As DeepHR calibrates the heart rate measurements that are mostly corrupted by motion, DeepHR clearly struggles when the presence of the motion and the original HR error are much more trivial. The second reason why DeepHR fails is the drop of heart rate measurements on Gear described in Section 5.1 that covers the relationship between motion and the HR error. The dropped periods are usually when motion is observed and results in HR errors. DeepHR would otherwise learn the relationship between motion and HR error from these period if the HR data was available instead of being dropped.

7.5 Summary

In this section, we evaluated DeepHR comprehensively with both controlled user study data and uncontrolled everyday data to demonstrate its potential in improving the noisy PPG-based heart rate measurements on the wearable. According to our results, DeepHR can indeed improve the heart rate measurement accuracy in everyday use across diverse activities. DeepHR also is capable of generalizing the predictions to unseen target activities and can be fine-tuned to improve the prediction performance with small amount of data from target activity incorporated in the training data. Besides, DeepHR is benchmarked against other two deep learning model that rely on conventional feature engineering to show the advantages brought by the convolutional and recurrent modules. The effective feature extraction by

convolutional module and the temporal dependencies captured by recurrent modules are crucial for DeepHR to be robust and able to generalize well. However, DeepHR struggles when the overall HR error is small and the motion level remains low as the situation found in everyday data collection with Samsung Gear devices. In next section, we discuss some topics related to our work as well as the future works and improvements that can be done on top of the thesis work.

8 Discussion

We first discuss the well-being monitoring that is based on HR monitoring and reference sensor for HR monitoring. Next, we discuss application areas, limitations, and future works of our study.

Well-being Monitoring Previous works[13, 14, 55] have explored approaches to monitor the well-being condition of the user with sensors equipped on the wearable. As we have shown, the accuracy of wrist-worn HR monitors for identifying the health related issues is still far away from satisfying. While the accuracy is not suitable for accurate clinical monitoring[9], there are other uses, like exercise intensity monitoring. Even with the limited capability in accurate continuous heart rate monitoring, the current wearable can detect anomalies as a way to warn abnormal health conditions[10]. The anomaly detection operates by opportunistically heart rate sampling, potentially integrating motion detection to avoid measurements when clear motion artifacts are present. The detected anomalies cannot lead to a direct diagnosis of the well-being issue with sufficient evidence due to the inaccurate heart rate monitoring, but can be used to hint that a further medical check is needed. Even the performance of heart rate monitoring improves as the wearable devices develops, which is illustrated by comparing the performance of Microsoft Band and Samsung Gear S3, it has not yet reached a reliable level where medical diagnosis can be based on. Our work offers a new insight into how to make the heart rate monitoring more reliable for medical and other use.

[9] https://www.independent.co.uk/life-style/health-and-families/health-news/heart-rate-monitor-unreliable-fitbit-garmin-health-science-technology-a8413196.html

[10] https://www.techradar.com/news/the-doctor-on-your-wrist-how-wearables-are-revolutionizing-healthcare

Reference Sensor The Polar H7 heart rate monitoring was chosen as the source of reference heart rate information instead of using medical level monitoring devices. The chest-worn heart rate belt offers ECG-based heart rate that is much more accurate than the PPG-based heart rate on the wrist-type wearable, and is low-cost and easily available compared to the medical heart rate monitoring devices, and most importantly it can be used to collect data from different kinds of daily activities unobtrusively without influencing the performance of normal routines. According to previous studies [92, 45], the Polar H7 HR monitoring strap has very high correspondence with medical level HR monitoring device. Considering its accessibility, unobtrusiveness, and high correspondence with the medical level device, the Polar H7 is adequate to be the reference in our work. Note that although we utilize data collected from Polar H7, DeepHR is not constrained to a certain types of heart rate monitoring devices as reference sensor, and instead it can utilize any kinds of high accuracy device as heart rate monitoring reference.

Application Area while DeepHR is applied to calibrate the heart rate monitoring on the wearable, the whole processing pipeline of DeepHR can be applied to other sensing and calibration applications. For instances, the deep learning model in DeepHR can be adapted for calibrating the air quality monitoring [70] and for calibrating a thermal camera [63]. In addition, the technique of motion representation and feature extraction with convolutional module can be utilized in scenarios like sleep monitoring [28] and transportation planning [29, 74].

Improvement on DeepHR DeepHR can improve the heart rate monitoring accuracy in both everyday use and user study. While the overall error after calibration in everyday use is reasonable (6.97 bpm), the model struggles for activities of high intensity in user study. This can be further improved by collecting more activity data which has the similar characteristics in the training data to fine-tune the model as we have shown. Alternatively, a dedicated model can be trained for a particular activity to achieve better calibration performance while sacrificing some generalizability of the model. For example, DeepHR could be trained with massive data collected during running activity and be used for calibrating the heart rate monitoring during running exercise, which is crucial to fitness training and coaching.

Energy Efficiency Running computationally intensive deep learning on the wearable can easily raise energy consumption overhead. In practice, DeepHR can be

pre-trained on a computationally rich device and then be deployed onto the wearable. As the wearable is usually used simultaneously with mobile phones, the computation can also be offloaded to the phone when possible [22]. This can reduce the energy consumption and accelerate the computation. In addition, deep learning compression techniques [31] and techniques for accelerating deep learning on the wearable [52] can be potentially applied to mitigate the computational and energy overhead caused by the deep learning model.

Federated Learning Federated learning is a machine learning technique that trains a model across decentralized devices. Each device holds data locally and learning operates without gathering or exchanging the local data of each client [46, 65, 47]. As we have shown in Section 7.3 that even small amount of training data collected from the target activity can improve the prediction performance of DeepHR, it can benefit from federated learning since information of a wider range of activities is present in the training data. This technique enables maximizing the utilization of user generated local data without compromising user data privacy. Federated learning operates by first training the customized local model on the devices and then aggregates incrementally learned model from the local clients into a global master model, which can be distributed to both existing and new clients in the system. With the development of IoT, more sensors and devices are connected together and data collected from multiple sensors can be shared with each other within a small area, such as in a smart home. Collected sensing data can be transmitted to and processed at a more powerful device, like a smartphone, a laptop, or a smart TV. These capabilities resulting from the development of IoT technology offers a promising application domain for federated learning. DeepHR can benefit from federated learning as both the wrist-type wearable and the heart rate belt are collected together in a IoT network, where the heart rate data (both PPG and ECG) and the motion sensing information (accelerometer and gyroscope) can be collected. Then the customized local DeepHR models can be trained to offer better prediction accuracy for the specific users and meanwhile the increments from local clients can be aggregated into a master model. How the local models should be integrated into the master model and how the master model should be used to update the local models to achieve the optimal performance is a possible future research direction.

9 Conclusion

As heart rate monitoring on wrist-worn wearable is becoming increasingly popular, this thesis work focuses on evaluating the heart rate monitoring performance of the off-the-shelf wearable devices, analysis of the HR measurement errors with respect to motion and other sources of error, and proposing a DeepHR as an approach to calibrate the inaccurate heart rate measurement due to motion and other errors.

Heart Rate Monitoring in Everyday Use We carried out a comprehensive user study that covers 9 different typical types of activities in daily life to evaluate the performance of the wearable heart rate monitoring devices. Previous studies have suggested the PPG-based heart rate monitoring provided by wearable can offer high accuracy and good correspondence with reference heart rate. However, most of the studies have been limited to rest activities or walking/running activities in lab conditions. Our work extends previous studies by assessing the performance of the wearable under everyday use. The evaluation shows that while the wearable can provide good accuracy for rest activities, the performance degrades notably for activities where different levels and types of motion are involved. The poor performance of continuous daily heart rate monitoring on the wearable devices constraints its application in psychological, physiological, and health related area.

Analysis of Heart Rate Errors We analyzed heart rate measurement performance thoroughly with respect to motion information. Overall, the HR error gets higher as the motion intensity increases, while the subtle and intermediate hand and wrist motion, such as while playing Rubik's cube and folding clothes, can also cause considerable errors. To quantify the motion information, motion indexes *acc_var* and *gyr_var* were built using the accelerometer and gyroscope measurements. The *acc_var* can indicate the movement of the hand while the *gyr_var* can indicate the rotation of the hand. Even though it is known the HR error is largely due to motion, the actual relationship between the motion and the HR error remains complex as also shown by our analysis.

Calibration Model To calibrate the motion induced HR errors, we proposed DeepHR, a deep learning model and processing pipeline, that captures the relationship between motion and error. First, motion representations are built from the accelerometer and gyroscope measurements. Then, a convolutional module is

applied to extract features from the motion representation, followed by a recurrent module to process temporal dependencies of the time-series data, and finally MLP layers to fine-tune the final output as the estimated HR error. We evaluate DeepHR with extensive data from both uncontrolled everyday settings and controlled user study to show its good performance and generalizability across different users and activities. Overall, DeepHR can effectively decrease the motion induced heart rate errors.

This thesis focuses on the optical heart rate monitoring on the wrist-worn wearable. We have shown that HR monitoring on the wearable devices is highly susceptible to motion in everyday use. This offers insights into the quality of continuous HR monitoring on the wearable devices for other applications built on top of it. We propose DeepHR as an approach to mitigate the HR errors utilizing motion information. Our evaluation demonstrates the good performance of DeepHR and suggests its potential for reducing errors in noisy sensor measurements.

References

1 J. Achten and A. E. Jeukendrup, "Heart rate monitoring," *Sports medicine*, vol. 33, pp. 517–538, 2003.

2 S. R. Aguilar, J. L. M. Merino, A. M. Sánchez, and Á. Á. Sánchez, "Variation of the heartbeat and activity as an indicator of drowsiness at the wheel using a smartwatch," *IJIMAI*, vol. 3, pp. 96–100, 2015.

3 J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, p. R1, 2007.

4 B. Askarian, K. Jung, and J. W. Chong, "Monitoring of heart rate from photoplethysmographic signals using a samsung galaxy note8 in underwater environments," *Sensors*, vol. 19, p. 2846, 2019.

5 L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, 2004, pp. 1–17.

6 S. Bhattacharya, H. Blunck, M. B. Kjærgaard, and P. Nurmi, "Robust and energy-efficient trajectory tracking for mobile devices," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 430–443, 2014.

7 S. Bhattacharya, H. Blunck, M. Kjærgaard, and P. Nurmi, "Robust and energy-efficient trajectory tracking for mobile devices," *IEEE Transactions on Mobile Computing*, vol. 14, p. 2, 2015.

8 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

9 A. J. Casson, A. V. Galvez, and D. Jarchi, "Gyroscope vs. accelerometer measurements of motion from wrist ppg during physical exercise," *ICT Express*, vol. 2, pp. 175 – 179, 2016.

10 D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *International journal of biosensors & bioelectronics*, vol. 4, p. 195, 2018.

11 Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1488–1492.

12 Y.-H. Chen, H.-H. Chen, T.-C. Chen, and L.-G. Chen, "Robust heart rate measurement with phonocardiogram by on-line template extraction and matching," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1957–1960.

13 J. Choi and R. Gutierrez-Osuna, "Using heart rate monitors to detect mental stress," in *Proceedings of the Sixth International Workshop on Wearable and Implantable Body Sensor Networks (BodyNets)*, 2009, pp. 219–223.

14 N. S. Chudy, "Testing of wrist-worn-fitness-tracking devices during cognitive stress: A validation study," Master's thesis, University of Central Florida, 2017.

15 W. G. Cochran and G. M. Cox, "Experimental designs. john willey and sons," *Inc., New York*, pp. 546–568, 1957.

16 E. Cope, "Estimating human movement using a three axis accelerometer," Ph.D. dissertation, Arizona State University, 2009.

17 S. E. Crouter, J. R. Churilla, and D. R. Bassett, "Estimating energy expenditure using accelerometers," *European journal of applied physiology*, vol. 98, no. 6, pp. 601–612, 2006.

18 C. J. Dondzila, C. Lewis, J. R. Lopez, and T. Parker, "Congruent accuracy of wrist-worn activity trackers during controlled and free-living conditions," *International Journal of Exercise Science*, vol. 11, pp. 575–584, 2018.

19 F. El-Amrawy, B. Pharm, and I. Nounou, "Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?" *Healthcare Informatics Research*, vol. 21, pp. 315–320, 2015.

20 B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," *Journal of Clinical Monitoring and Computing*, vol. 27, pp. 313–317, 2013.

21 D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645–662, 2010.

22 H. Flores, P. Hui, P. Nurmi, E. Lagerspetz, S. Tarkoma, J. Manner, V. Kostakos, Y. Li, and X. Su, "Evidence-aware mobile computational offloading," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 1834–1850, 2017.

23 T. K. Fredericks, S. D. Choi, J. Hart, S. E. Butt, and A. Mital, "An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads," *International Journal of Industrial Ergonomics*, vol. 35, pp. 1097 – 1107, 2005.

24 R. Gonzalez, A. Manzo, J. Delgado, J. Padilla, B. Trénor, and J. Saiz, "A computer based photoplethysmographic vascular analyzer through derivatives," in *2008 Computers in Cardiology*. IEEE, 2008, pp. 177–180.

25 I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

26 A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

27 S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

28 W. Gu, L. Shangguan, Z. Yang, and Y. Liu, "Sleep hunter: Towards fine grained sleep stage tracking with smartphones," *IEEE Transactions on Mobile Computing*, vol. 15, pp. 1514–1527, 2015.

29 W. Gu, M. Jin, Z. Zhou, C. J. Spanos, and L. Zhang, "Metroeye: smart tracking your metro trips underground," in *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2016, pp. 84–93.

30 N. Y. Hammerla and T. Plötz, "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 1041–1051.

31 S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

32 K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

33 S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *Proceedings of the 11th ACM conference on embedded networked sensor systems*, 2013, p. 13.

34 ——, "Gravity and linear acceleration estimation on mobile devices," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2014, pp. 50–59.

35 J. Hernandez, D. McDuff, K. Quigley, P. Maes, and R. W. Picard, "Wearable motion-based heart rate at rest: A workplace evaluation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 1920–1927, 2019.

36 A. B. Hertzman, "The blood supply of various skin areas as estimated by the photoelectric plethysmograph," *American Journal of Physiology*, vol. 124, pp. 328–340, 1938.

37 S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, pp. 107–116, 1998.

38 M. Kaisti, T. Panula, J. Leppänen, R. Punkkinen, M. J. Tadi, T. Vasankari, S. Jaakkola, T. Kiviniemi, J. Airaksinen, P. Kostiainen *et al.*, "Clinical assessment of a non-invasive wearable mems pressure sensor array for monitoring of arterial pulse waveform, heart rate and detection of atrial fibrillation," *NPJ digital medicine*, vol. 2, pp. 1–10, 2019.

39 F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

40 J. Karvonen and T. Vuorimaa, "Heart rate and exercise intensity during sports activities," *Sports medicine*, vol. 5, pp. 303–311, 1988.

41 A. Khushhal, S. Nichols, W. Evans, D. O. Gleadall-Siddall, R. Page, A. F. O'Doherty, S. Carroll, L. Ingle, and G. Abt, "Validity and reliability of the apple watch for measuring heart rate during exercise," *Sports Medicine International Open*, vol. 1, no. 06, pp. E206–E211, 2017.

42 B. S. Kim and S. K. Yoo, "Motion artifact reduction in photoplethysmography using independent component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 566 – 568, 2006.

43 D. H. Kim, Y. Kim, D. Estrin, and M. B. Srivastava, "Sensloc: sensing everyday places and paths using less energy," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2010, pp. 43–56.

44 K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and Biological Engineering and Computing*, vol. 42, pp. 419–427, 2004.

45 M. Kingsley, M. J. Lewis, and R. Marson, "Comparison of polar 810 s and an ambulatory ecg system for rr interval measurement during progressive exercise," *International journal of sports medicine*, vol. 26, pp. 39–44, 2005.

46 J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

47 J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

48 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

49 J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, pp. 74–82, 2011.

50 M. T. La Rovere, G. D. Pinna, R. Maestri, A. Mortara, S. Capomolla, O. Febo, R. Ferrari, M. Franchini, M. Gnemmi, C. Opasich *et al.*, "Short-term heart rate variability strongly predicts sudden cardiac death in chronic heart failure patients," *circulation*, vol. 107, pp. 565–570, 2003.

51 N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 283–294.

52 N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Qendro, and F. Kawsar, "Deepx: A software accelerator for low-power deep learning inference on mobile devices," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, 2016, p. 23.

53 A. Leatham, "Phonocardiography," *British medical bulletin*, vol. 8, pp. 333–342, 1952.

54 Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, p. 436, 2015.

55 B.-G. Lee, B.-L. Lee, and W.-Y. Chung, "Smartwatch-based driver alertness monitoring with wearable motion and physiological sensor," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 6126–6129.

56 C. M. Lee and M. Gorelick, "Validity of the smarthealth watch to measure heart rate during rest and exercise," *Measurement in Physical Education and Exercise Science*, vol. 15, no. 1, pp. 18–25, 2011.

57 H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

58 C. D. B. Luft, E. Takase, and D. Darby, "Heart rate variability and cognitive function: Effects of physical effort," *Biological psychology*, vol. 82, pp. 186–191, 2009.

59 Y. Maeda, M. Sekine, and T. Tamura, "Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography," *Journal of Medical Systems*, vol. 35, pp. 969–976, 2011.

60 ——, "The advantages of wearable green reflected photoplethysmography," *Journal of Medical Systems*, vol. 35, pp. 829–834, 2011.

61 R. Mahony, T. Hamel, and J.-M. Pflimlin, "Nonlinear complementary filters on the special orthogonal group," *IEEE Transactions on automatic control*, vol. 53, pp. 1203–1218, 2008.

62 M. Malik, "Heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 1, pp. 151–181, 1996.

63 T. Malmivirta, J. Hamberg, E. Lagerspetz, X. Li, E. Peltonen, H. Flores, and P. Nurmi, "Hot or not? robust and accurate continuous thermal imaging on flir cameras," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, 2019, pp. 1–9.

64 D. McDuff, S. Gontarek, and R. W. Picard, "Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2948–2954, 2014.

65 H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

66 T. Michel, P. Geneves, H. Fourati, and N. Layaïda, "On attitude estimation with smartphones," in *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2017, pp. 267–275.

67 R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, pp. 1236–1246, 2017.

68 D. M. Mirvis and A. L. Goldberger, "Electrocardiography," *Heart disease*, vol. 1, pp. 82–128, 2001.

69 A. V. Moço, S. Stuijk, and G. de Haan, "Skin inhomogeneity as a source of error in remote ppg-imaging," *Biomedical optics express*, vol. 7, pp. 4718–4733, 2016.

70 N. H. Motlagh, E. Lagerspetz, P. Nurmi, X. Li, S. Varjonen, J. Mineraud, M. Siekkinen, A. Rebeiro-Hargrave, T. Hussein, T. Petaja, M. Kulmala, and S. Tarkoma, "Toward massive scale air quality monitoring," *IEEE Communications Magazine*, vol. 58, pp. 54–59, 2020.

71 C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

72 F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, p. 115, 2016.

73 D. Phan, L. Y. Siong, P. N. Pathirana, and A. Seneviratne, "Smartwatch: Performance evaluation for long-term heart rate monitoring," in *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)*. IEEE, 2015, pp. 144–147.

74 A. C. Prelipcean, G. Gidofalvi, and Y. O. Susilo, "Measures of transport mode segmentation of trajectories," *International Journal of Geographical Information Science*, vol. 30, pp. 1763–1784, 2016.

75 M. Raghuram, K. V. Madhav, E. H. Krishna, N. R. Komalla, K. Sivani, and K. A. Reddy, "Dual-tree complex wavelet transform for motion artifact reduction of PPG signals," in *Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MEMA)*, 2012.

76 M. R. Ram, K. V. Madhav, E. H. Krishna, N. R. Komalla, and K. A. Reddy, "A Novel Approach for Motion Artifact Reduction in PPG Signals Based on AS-LMS Adaptive Filter," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 1445–1457, 2012.

77 ——, "A novel approach for motion artifact reduction in ppg signals based on as-lms adaptive filter," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 1445–1457, 2011.

78 R. K. Reddy, R. Pooni, D. P. Zaharieva, B. Senf, J. El Youssef, E. Dassau, F. J. Doyle III, M. A. Clements, M. R. Rickels, S. R. Patton *et al.*, "Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: Evaluation study," *JMIR mHealth and uHealth*, vol. 6, p. e10338, 2018.

79 C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235 – 244, 2016.

80 N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak, and S. Anand, "Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography," *Journal of medical engineering & technology*, vol. 32, pp. 479–484, 2008.

81 A. Shcherbina, C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler, and E. A. Ashley, "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," *Journal of Personalized Medicine*, vol. 7, 2017.

82 ——, "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," *Journal of personalized medicine*, vol. 7, p. 3, 2017.

83 D. K. Spierer, Z. Rosen, L. L. Litman, and K. Fujii, "Validation of photoplethysmography as a method to detect heart rate during rest and exercise," *Journal of medical engineering & technology*, vol. 39, pp. 264–271, 2015.

84 S. E. Stahl, H.-S. An, D. M. Dinkel, J. M. Noble, and J.-M. Lee, "How accurate are the wrist-based heart rate monitors during walking and running activities? are they accurate enough?" *BMJ open sport & exercise medicine*, vol. 2, p. e000106, 2016.

85 Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

86 T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors?past and present," *Electronics*, vol. 3, pp. 282–302, 2014.

87 X. Teng and Y.-T. Zhang, "The effect of contacting force on photoplethysmographic signals," *Physiological measurement*, vol. 25, no. 5, p. 1323, 2004.

88 M. Valentini and G. Parati, "Variables influencing heart rate," *Progress in cardiovascular diseases*, vol. 52, pp. 11–19, 2009.

89 A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

90 M. P. Wallen, S. R. Gomersall, S. E. Keating, U. Wisløff, and J. S. Coombes, "Accuracy of heart rate watches: Implications for weight management," *PLOS ONE*, vol. 11, pp. 1–11, 2016.

91 C. Wang, T. Pun, and G. Chanel, "A comparative survey of methods for remote heart rate detection from frontal face videos," *Frontiers in bioengineering and biotechnology*, vol. 6, p. 33, 2018.

92 R. Wang, G. Blackburn, M. Desai, D. Phelan, L. Gillinov, P. Houghtaling, and M. Gillinov, "Accuracy of wrist-worn heart rate monitors," *Jama cardiology*, vol. 2, pp. 104–106, 2017.

93 M. Woodward, R. Webster, Y. Murakami, F. Barzi, T.-H. Lam, X. Fang, I. Suh, G. D. Batty, R. Huxley, and A. Rodgers, "The association between resting heart rate, cardiovascular disease and mortality: evidence from 112,680 men and women in 12 cohorts," *European journal of preventive cardiology*, vol. 21, pp. 719–726, 2014.

94 J. Xie, D. Wen, L. Liang, Y. Jia, L. Gao, and J. Lei, "Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: comparative study," *JMIR mHealth and uHealth*, vol. 6, p. e94, 2018.

95 C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, pp. 7772–7788, 2010.

96 J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 3995–4001.

97 R. Yang, E. Shin, M. W. Newman, and M. S. Ackerman, "When fitness trackers don't 'fit': End-user difficulties in the assessment of personal tracking device accuracy," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015.

98 S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 351–360.

99 Y. Ye, Y. Cheng, W. He, M. Hou, and Z. Zhang, "Combining nonlinear adaptive filtering and signal decomposition for motion artifact removal in wearable photoplethysmography," *IEEE Sensors Journal*, vol. 16, pp. 7133–7141, 2016.

100 R. Yousefi, M. Nourani, S. Ostadabbas, and I. M. S. Panahi, "A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors," *IEEE Journal on Biomedical and Health Informatics*, vol. 18, pp. 670 – 681, 2014.

101 Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, pp. 522 – 531, 2015.