Passive and Context-Aware In-Home Vital Signs Monitoring Using Co-Located UWB-Depth Sensor Fusion

ZONGXING XIE, Stony Brook University, USA
BING ZHOU, IBM Thomas J. Watson Research Center, USA
XI CHENG, Stony Brook University, USA
ELINOR SCHOENFELD, Stony Brook University, USA
FAN YE, Stony Brook University, USA

Basic vital signs such as heart and respiratory rates (HR and RR) are essential bio-indicators. Their longitudinal in-home collection enables prediction and detection of disease onset and change, providing for earlier health intervention. In this paper, we propose a robust, non-touch vital signs monitoring system using a pair of co-located Ultra-Wide Band (UWB) and depth sensors. By extensive manual examination, we identify four typical temporal and spectral signal patterns and their suitable vital signs estimators. We devise a probabilistic weighted framework (PWF) that quantifies evidence of these patterns to update the weighted combination of estimator output to track the vital signs robustly. We also design a "heatmap" based signal quality detector to exclude the disturbed signal from inadvertent motions. To monitor multiple co-habiting subjects in-home, we build a two-branch long short-term memory (LSTM) neural network to distinguish between individuals and their activities, providing activity context crucial to disambiguating critical from normal vital sign variability. To achieve reliable context annotation, we carefully devise the feature set of the consecutive skeletal poses from the depth data, and develop a probabilistic tracking model to tackle non-line-of-sight (NLOS) cases. Our experimental results demonstrate the robustness and superior performance of the individual modules as well as the end-to-end system for passive and context-aware vital signs monitoring.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Applied computing → Health informatics**; **Bioinformatics**.

Additional Key Words and Phrases: Non-touch vital signs monitoring; passive sensing, longitudinal in-home monitoring

1 INTRODUCTION

Basic vital signs including respiration and heart rates are predictors for assessing overall changes in health status [11], and a myriad of medical conditions including respiratory, cardiac, and sleep conditions [45, 51].

Continuous vital signs data collected in individuals' home environment can be analyzed to monitor disease onset/progression/resolution, and the impact of new or changed medications. Such in home assessment can have tremendous benefits for anyone living with a chronic health condition, especially for older adults who face a myriad of chronic diseases and health conditions

Longitudinal in-home monitoring requires low-cost, robust and passive sensing. Traditional hospital equipment such as electrocardiograms (EKG) are expensive, are not designed for continuous in-home data collection, and require well-trained medical personnel to set up and monitor the output. Despite their popularity, wearables

Part of this work was presented at the 9th IEEE International Conference on Healthcare Informatics (IEEE ICHI 2021), Victoria, Canada, August, 2021 [85]. This work is supported in part by NSF grants 1652276, 1951880, 2028952, 2119299. This work was conducted when Bing Zhou was with Stony Brook University.

Authors' addresses: Zongxing Xie, Stony Brook University, Department of Electrical and Computer Engineering, Stony Brook, NY, USA, zongxing.xie@stonybrook.edu; Bing Zhou, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, bing.zhou@ibm.com; Xi Cheng, Stony Brook University, Department of Electrical and Computer Engineering, Stony Brook, NY, USA, cheng.xi@stonybrook.edu; Elinor Schoenfeld, Stony Brook University, Department of Family, Population and Preventive Medicine, Stony Brook, NY, USA, elinor.schoenfeld@stonybrookmedicine.edu; Fan Ye, Stony Brook University, Department of Electrical and Computer Engineering, Stony Brook, NY, USA, fan.ye@stonybrook.edu.

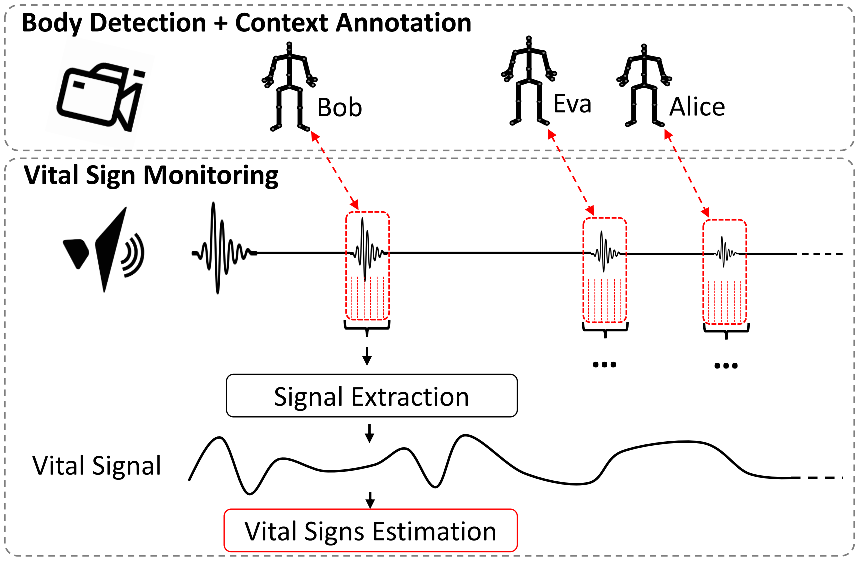


Fig. 1. *VitalHub* leverages depth camera for body detection and context annotation. The location of detected human body is used to segment the UWB signal correspodning to the chest wall of the interested subject, and vital signal is extracted for vital signs estimation.

(e.g., Apple Watch, Fitbit) have inconvenient and restraining daily maintenance overheads (e.g., charge, wear), especially difficult among physically and cognitively challenged older adults.

Recent radio-based passive-sensing solutions [29] exemplified by Wi-Fi [45, 58, 79], FMCW [5], and UWB [53, 64] hold promise for longitudinal in-home monitoring. Temporal and spectral methods extracting vital signs against multipath [79], cluttered environments [39] are proposed.

However, robustness against harmonics and intermodulation has not received sufficient attention. Because neither heartbeat nor respiratory signals are purely sinusoidal and the phase of RF signals for vital signs extraction exhibits non-linearity, high-order harmonics and intermodulations (i.e., linear combinations of heart and respiration rates) exist, and frequently carry energy stronger than the fundamental heart rate. They produce high spectral peaks within the normal heart rate range, e.g., 50−150 beat per minute (bpm). Thus, spectral methods simply pick the highest peaks to identify the heartbeat signal easily fail (1/3 of the time in our experiments). Making matters worse, we observe that their frequencies and magnitudes are time-varying, and the pattern keeps changing over time, defying simple predictions. These issues are not described nor appear tackled in recent radio sensing work. The electrical engineering community has conducted some related studies [21, 64], but in-depth validation and conclusive comparison are still lacking.

Similarly, robustness against signal corruption remains elusive. Due to inevitable large body motion, the signal might be corrupted beyond recognition by even well-trained humans. Such signals must be detected and excluded to avoid producing erroneous results. Existing methods [5, 29] rely on spectral energy or temporal waveform assumptions that are susceptible to dynamic changes, thus not reliable enough.

Longitudinal in-home monitoring also needs to address two other issues: 1) differentiate multiple co-habiting subjects in proximity, reliably extract and associate vital signs data to each of them; 2) identify the physical activity context to disambiguate pathological and normal changes in vital signs (e.g., the heart rate increases after exercises or food is normal but during sleep apnea [45, 51] is abnormal).

In this paper, we propose *VitalHub* (as shown in Figure 1), a robust vital sign monitoring system meeting the requirements of longitudinal in-home monitoring using a pair of co-located UWB and depth sensors. Based on a manual examination of over 6000 data samples, we identify four typical temporal and spectral patterns (present in 98.55% of the data) and a suitable heart/respiration rate estimator for each. To handle harmonics and intermodulation, we devise a probabilistic weighted framework (PWF) that quantifies the cumulative evidence of these patterns to adaptively update the weighted combination of estimator outputs to track the vital signs robustly. To detect corrupted signals, we generate a 2D "heatmap" representing the likelihood of different heart/respiration rate estimates and train a ResNet [28] model to produce a confidence value of how likely the signal is corrupted.

The depth sensor provides multiple functions: 1) recognize human bodies and their relative depths, which enable us to identify which "range bins" (i.e., signals reflected from objects within certain depth ranges) correspond to the human bodies for the segmentation and extraction of vital signals as illustrated in Figure 1; 2) differentiate co-habiting subjects from skeletal walking patterns [70] for proper data association; and 3) recognize body poses to produce context of physical activities. Unlike RGB cameras, the depth sensor does not have fine visual details, thus more privacy-friendly to subjects.

We have implemented a VitalHub prototype and conducted extensive experiments. We collected data from 8 volunteers in 56 sessions (2–10 *min* per session) in both stationary (e.g., sitting still) and non-stationary (e.g., natural upper body swaying) poses at 3 different distances/angles. We spent over 72 man-hours to manually label more than 40,000 30s time-windowed signals whether they were corrupted beyond human recognition to provide training data.

Our PWF aided by the detector achieves 1.5/3.2 bpm error at 80-percentile for respiration and heart rates, even though individual estimators may produce 10–20 bpm errors in heart rates. [1] These are very close to 1.2/1.5 bpm errors from an idealistic *oracle* that always knows whether the signal is corrupted, which are the best instantaneous range bin and estimator (none of which practically feasible).

We make the following contributions in this work:

- We systematically describe the challenges of non-touch vital signs monitoring using a low-cost COTS UWB sensor in realistic in-home scenarios. The challenges manifest in the form of dynamic signal patterns with harmonics and intermodulation interference, making robust heart rate estimation difficult, for which we do not find sufficient description nor treatment in the literature.
- We design a probabilistic weighted framework (PWF) that adaptively adjusts the weights in combining the outputs of four estimators based on quantitative evidence of respective patterns, to address the challenges caused by the harmonics and intermodulation interference for heart rate estimation. Extensive experiments show that our proposed system achieves within 0.3/1.7 bpm error to the upper limit of an idealistic oracle, demonstrating the robustness of PWF.
- To better understand the contributions of our vital signs monitoring pipeline, we conduct comparative evaluation with related work. Specifically, we compare VitalHub with three representative methods dealing with harmonics and intermodulation issues. We show that VitalHub achieves ≤5bpm error 98.5% of the time in heart rate estimation, while others only 51.2–81.3%, and we share insights how assumptions they rely on may not hold in reality. We also compare our heatmap based signal quality detector against other 4 common methods, and find it achieves near-human performance at 96% for both precision and recall in detecting and excluding corrupted signals caused by inadvertent motions, while others at best 89/83%.
- We develop a context annotation module, including a two-branch LSTM recurrent neural network and a probabilistic model, for identity tracking and activity recognition with a feature set derived from the skeleton data, using a depth camera independent of the ambient lighting conditions. Experiments show

[1] The respiration rate is more accurate due to the stronger energy.

that the LSTM model outperforms baseline classifiers and achieves 90% median accuracy for differentiating 8 people from skeletal walking patterns, and 96% F1-score for recognizing 6 common daily activities.

VitalHub produces robust measurements for respiration and heart rates against harmonics and intermodulation. It reliably differentiates co-habiting subjects to associate data, and generates activity contexts, thus offering a suitable solution for longitudinal in-home vital signs data collection valuable for future customized health analytics (e.g., the detection of anomalous deviation from a user's normal patterns of vital signs during certain daily activities).

2 DESIGN CONSIDERATIONS

2.1 Design Goals

To achieve longitudinal in-home vital signs monitoring, we identify several goals as follows:

- **Robustness.** Vital signs should be robustly extracted against dynamic changes including corruptions, strong harmonics and intermodulation, and in the presence of multiple co-habiting people.
- **Passive, non-touch sensing.** The monitoring should not require active user efforts such as charging batteries or wearing devices. This is critical for older adults who would benefit the most from longitudinal monitoring but many are physically and cognitively challenged for active efforts.
- **Context awareness.** Co-habiting subjects must be distinguished from each other so as to properly associate the data to the corresponding individuals. The activity context of the subjects is needed to help disambiguate abnormal changes (physical or mental) from normal ones.
- **Privacy-friendly.** We have conducted 5 discussion groups, each 10-15 participants (a total of 57 older adults) about their perception on technologies for in-home health monitoring. They come from both urban and suburban areas, with diversity in gender, age (60 to 80+), and socioeconomic status. Each time the participants were asked to self complete a paper-based questionnaire at the beginning, so their opinions were not influenced or biased by our later presentation. In questionnaires, we found people did not want to be monitored by regular RGB cameras which show fine grained images/videos. In discussions they overwhelmingly expressed strong disinterests and privacy concerns, no matter who is watching, even close family members like adult children living separately. However, they were more receptive to coarse-grained silhouettes from depth sensors because no visual details of facial expression, clothing were available.

2.2 Hardware Choices

To achieve the above goals, we choose a co-located *UWB* and *depth* sensor pair as the hardware platform. The UWB signal is sensitive to tiny displacements of the chest wall due to the heartbeat and respiration. The UWB system is found to be highly immune to multipath effects with energy spanning a wide frequency bandwidth [40], and requires less complexity in the architecture than FMCW ones [8, 76] for wireless sensing. [2] A less complexity in the architecture usually implies a less expensive COTS solution for low-cost deployment, thus we choose a UWB sensor [7] as the RF frontend for non-touch vital signs monitoring.

To balance the requirements of context-awareness and privacy-friendliness, we adopt a depth camera to 1) detect and locate human bodies, so as to help select the range bins in the received UWB signals corresponding to the chest walls; 2) distinguish multiple people simultaneously present in the Field-of-View (FoV) for data-identity association; and 3) produce context information without using intrusive RGB images. Notably, the depth camera in our implementation uses IR based time-of-flight method for depth sensing, and the human body detection [70]

[2] To produce the range profile, the UWB based system needs only a down-conversion mixer; In contrast, the FMCW based system needs 1D-FFT in addition to the down-conversion mixer, because its time of flight (TOF) has to be linearly translated from the frequency shift.

is based on the depth image without RGB data. Therefore, the features of VitalHub enabled by the depth camera are independent of the ambient lighting conditions, so it works well at both day and night.

2.3 Background of UWB based Vital Signs Extraction

This section provides necessary background regarding vital signs extraction via UWB signal modeling. Heart and respiratory rates are two of the five vital signs collected at each physical examination [11]. It is the combination of heartbeat and respiration that comprises chest displacements. The instantaneous distance ($d(t)$) of the chest wall away from the UWB sensor can be measured with high sensitivity for vital signs extraction, and can ideally be expressed [63] as:

$$\begin{aligned} d(t) &= d_0 + D(t) \\ &= d_0 + d_r \sin\left(2\pi f_r t\right) + d_h \sin\left(2\pi f_h t\right), \end{aligned} \tag{1}$$

where d_0 is the nominal distance between the UWB sensor and the targeted chest wall (i.e., provided by the depth sensor to select the proper "range bin"), $D(t)$ is the chest wall displacement; d_r and d_h are the displacement amplitudes, and f_r and f_h the rates of respiration and heartbeat, respectively.

The instantaneous channel response $h(t, \tau)$ at time t with a short delay τ can be formulated as:

$$h(t, \tau) = \alpha_D \delta\left(\tau - \tau_D(t)\right) + \sum_{i=1} \alpha_i \delta\left(\tau - \tau_i\right), \tag{2}$$

where α_D and α_i represent the magnitudes of the channel response of the target and other static objects, and $\tau_D(t) = 2d(t)/c$ and τ_i are corresponding delays (c the speed of light). It indicates that the channel response of the target can be spatially distinguished from the clutter according to the range/TOF. The received signal $s(t, \tau)$ can be derived as a convolution of the channel response and the transmitted impulse $p(\tau)$, as:

$$\begin{aligned} s(t, \tau) &= p(\tau) * h(t, \tau) \\ &= \alpha_D p\left(\tau - \tau_D(t)\right) + \sum_{i=1} \alpha_i p\left(\tau - \tau_i\right). \end{aligned} \tag{3}$$

Therefore, the segment of received signal $\alpha_D p\left(\tau - \tau_D(t)\right)$ shifts in the phase according to the two-way echo delay $\tau_D(t) = 2d(t)/c$ due to the chest displacement, and two vital signs (heart and respiratory rates) can be extracted from the phase. The phase can be modeled as:

$$\phi(t) = \phi_0 + \phi_D(t), \tag{4}$$

where ϕ_0 is the initial phase of the received signal at the nominal distance d_0, $\phi_D(t) = 2\kappa D(t)$ is the phase modulated by the physiological movements, and $\kappa = 2\pi/\lambda$ denotes the angular wavenumber, determined by the wavelength λ of the carrier wave.

2.4 Robustness Challenges

Robust vital sign extraction based on the derived phase model in (4) is challenging due to the following issues. First, the perceived phase can be noisy due to imperfect hardware. As the UWB signal is sampled at extremely high frequencies (23.328 GHz in our case), imperfect synchronization between the transmitter and receiver would result in a sampling time offset (STO), thus a time-variant phase drift $\phi_{STO}(t)$. Therefore, the phase model (4) needs to be updated as:

$$\phi(t) = \phi_0 + \phi_D(t) + \phi_{STO}(t), \tag{5}$$

and this makes direct extraction difficult especially when the phase drift from desynchronization becomes larger than that ($\phi_D(t)$) from physiological motion.

Second, the chest wall movements due to either heartbeat or respiration are not purely sinusoidal, thus harmonic components exist for both. As normal heart rates span a wide range (e.g., 50–150 bpm), the higher order harmonics of respiration can co-exist in the same range. Larger respiration motions also produce strong harmonics, making it difficult to decide the correct fundamental frequency of heart rate. To address this issue, we will introduce a probabilistic weighting framework (in §4.3.2), which adaptively reduces the interference of respiration harmonics to the heart rate estimation.

Third, in realistic scenarios, the phase of the received signal is more complex than linearly proportional to the chest wall displacement $\phi_D(t) = 2\kappa D(t)$. We start our reasoning with the scattering model of the human body for vital signs extraction [50, 82, 95]. The human body is more complex than a point scatterer. Rather, it has a 3D shape and should be modeled as a collection of point scatterers at different depths [50]. Therefore, the received signal is a superposition of the scattered signals from different body parts which may interfere with each other, constructively or destructively. If we categorize the scattered signals into two sets, one is modulated by physiological movements (denoted by M), and the other is static (denoted by N), the resulting signal can be expressed as:

$$\begin{aligned} s(t) &= \sum_{m \in M} \alpha_m e^{-j(\phi_m + 2\kappa D(t))} & &+ \sum_{n \in N} \alpha_n e^{-j\phi_n} \\ &= \alpha_{\bar{m}} e^{-j\phi_{\bar{m}}} e^{-j2\kappa D(t)} & &+ \alpha_{\bar{n}} e^{-j\phi_{\bar{n}}}, \end{aligned} \tag{6}$$

where the first term varies over time according to the physiological movements, and the second term indicates static, thus DC components. The subscripts m and n denote the elements from the set M and set N separately, $\bar{m}$ and $\bar{n}$ denote their resulting summed terms respectively, and ϕ denotes the phase offset from the relative distances between point scatterers. Then the phase of the resulting signal $s(t)$ can be obtained from the in-phase signal ($I(t) = \mathfrak{R}\{s(t)\}$) and quadrature signal ($Q(t) = \mathfrak{I}\{s(t)\}$) with the arctangent demodulation method:

$$\phi_D(t) = \arctan \frac{Q(t)}{I(t)} = \arctan \frac{\alpha_{\bar{m}} \cos(2\kappa D(t) + \phi_{\bar{m}}) + \alpha_{\bar{n}} \cos(\phi_{\bar{n}})}{\alpha_{\bar{m}} \sin(2\kappa D(t) + \phi_{\bar{m}}) + \alpha_{\bar{n}} \sin(\phi_{\bar{n}})}. \tag{7}$$

Therefore, the resulting phase of the received signal $s(t)$ can be expressed as a nonlinear function in terms of $D(t)$, which can be approximated by its Taylor series as follows:

$$\phi_D(t) = \sum_{i=1}^{\infty} a_i D^i(t) = (a_1 D(t) + a_2 D^2(t) + a_3 D^3(t) + \ldots), \tag{8}$$

where a_i is the coefficient of the i-th order term. The higher order terms in the Taylor series of the nonlinear signal result in the intermodulation products between heartbeat and respiratory signals [39]. Take the second order term for example:

$$D^2(t) = d_r^2 \sin(2\pi f_r t)^2 + d_h^2 \sin(2\pi f_h t)^2 + 2 d_r d_h \sin(2\pi f_r t) \sin(2\pi f_h t). \tag{9}$$

According to the product-to-sum formulas, the trigonometric functions in (9) can be expressed as:

$$\begin{aligned} \sin(2\pi f_r t)^2 &= \frac{1}{2} - \frac{1}{2} \cos(\underline{4\pi f_r t}), \\ \sin(2\pi f_h t)^2 &= \frac{1}{2} - \frac{1}{2} \cos(\underline{4\pi f_h t}), \\ \sin(2\pi f_h t) \sin(2\pi f_r t) &= \frac{1}{2} \cos(\underline{2\pi (f_h - f_r) t}) - \frac{1}{2} \cos(\underline{2\pi (f_h + f_r) t}). \end{aligned} \tag{10}$$

As indicated in the underlined items in (10), intermodulation components manifest as spectral components at frequencies which are linear combinations of heart and respiratory rates (i.e., $\{m f_h \pm n f_r | m, n \in \mathbb{N}_0\}$). Such components could exist in the normal heart rate range, making it even more difficult to determine the correct frequency component of heartbeat. The coefficients a_i in (8) are time-variant, resulting in unpredictable and

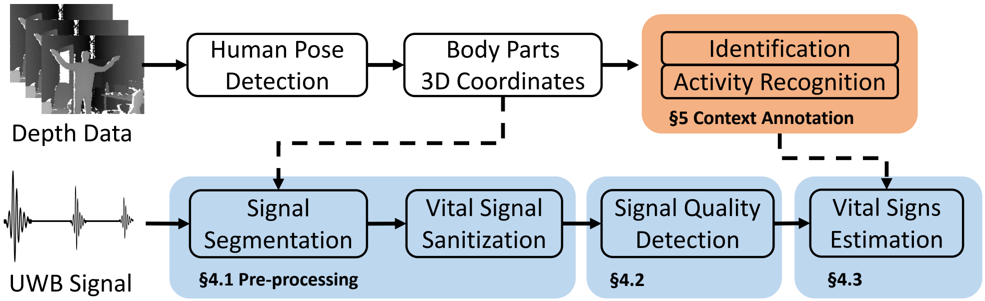


Fig. 2. The overall framework of VitalHub. Two parallel data streams from the UWB and depth sensors are complementary and combined for passive and context-aware vital signs monitoring. The pre-processing module segments vital signals out of the UWB signal reflected from the cluttered environment according to the corresponding location of the target chest wall from the depth data stream. The signal quality detection module suspends vital signs estimation upon corrupted signals caused by disturbances from random body movements. The vital signs estimation module addresses the robust challenges, and the vital signs data will be automatically annotated with the corresponding identification and activity context for future customized analytics.

dynamic magnitudes of intermodulation components. Therefore, a method that relies on certain assumptions on signal patterns may fail under different patterns.

3 VITALHUB OVERVIEW

Figure 2 illustrates the overall framework of VitalHub, which fuses inputs from a pair of co-located UWB and depth sensors to tackle challenges described in §2.4 for robust and context-aware vital signs monitoring. So there are two streams of data (i.e., UWB and depth data) processed in parallel, and combined organically.

To process the UWB data stream for vital signs extraction, we first develop a pre-processing pipeline (detailed in §4.1) to deal with STO issues and extract the vital signal (i.e., the phase change in UWB signal due to physiological motions). Then we introduce a signal quality detector (detailed in §4.2) to tell signals where vital signs are "available" for estimation from corrupted ones due to inadvertent movements, even in the presence of harmonics and intermodulation. Moreover, we propose a probabilistic weighted framework (PWF) in vital signs estimation (detailed in §4.3) to specifically deal with the challenges in robust heart rate estimation in the presence of dynamic signal patterns.

While the UWB sensor is sensitive to minute movements for vital signs estimation, it is relatively "blind" to the context information (e.g., where the subject of interest is located, which subjects are present, and what activities each subject is conducting). We leverage the complementary characteristics of the depth sensor to achieve context-awareness.

We process the depth data to support unambiguous monitoring in cohabiting scenarios to correctly associate respective context (e.g., identities and activities) to the UWB echo pulses from different subjects. To be specific, we leverage a human pose recognition model [70] to detect the body parts thus the poses of the subjects present in the FoV of the depth sensor. It outputs 3D positions of body joints as a representation of skeletal pose. We use the predicted position of torso to help locate the segments of UWB signals corresponding to the chest wall for vital signal extraction (explained in §4.1). We further leverage spatial and temporal features of consecutive skeletal poses to generate the context information (detailed in §5). Context is needed for the completeness of the passive and context-aware monitoring system.

4 VITAL SIGNS MONITORING

In this section, we describe the vital signs monitoring module, which consists of three stages: 1) signal pre-processing to extract vital signals from received noisy UWB echoes; 2) signal quality detector; and 3) vital sign estimation to robustly measure heart/respiration rates in presence of unpredictable and dynamic signal patterns.

4.1 UWB Signal Pre-processing

We design a UWB signal pre-processing pipeline to extract vital signals (i.e., phase changes due to physiological movements) from the reflected UWB pulses.

Signal Segmentation. This step locates the segments of received UWB signals corresponding to the target (i.e., chest walls). The pulses reflected from different distances are received at different arrival times. Thus we segment signals into *range bins* each corresponding to a different 5 *cm* depth range. [3] Our UWB sensor has a range of 10 *m*, leading to about 200 range bins. As illustrated in Figure 1, we leverage the human body distance measurement from context annotation module (in §5) to decide which range bin corresponds to which identified human body, thus further processing signals in those bins.

Vital Signal Sanitization. Next we remove the time-variant phase drift $\phi_{STO}(t)$ due to sampling time offset (STO) (analyzed in §2.4). Because $\phi_{STO}(t)$ is caused by unknown jitters in the sampling system, it is impossible to describe with a mathematical model. Fortunately, the same jitters exist in signals from all range bins, and the direct path (i.e., the signal received from the transmitter, without reflection from any object). The direct path signal can be expressed as $\phi_r(t) = \phi_0^r + \phi_{STO}(t)$, where ϕ_0^r is the inital phase of the direct path signal and is static. Therefore, we can simply use $\phi_r(t)$ as a reference to cancel out $\phi_{STO}(t)$ as follows to obtain sanitized vital signals in the form of relative phases:

$$\phi'(t) = \phi(t) - \phi_r(t) = \phi_D(t) + \phi_0 - \phi_0^r, \tag{11}$$

where ϕ_0 and ϕ_0^r are both static, and $\phi_D(t)$ is the phase modulated by the physiological movements from which we estimate vital signs.

4.2 Signal Quality Detector

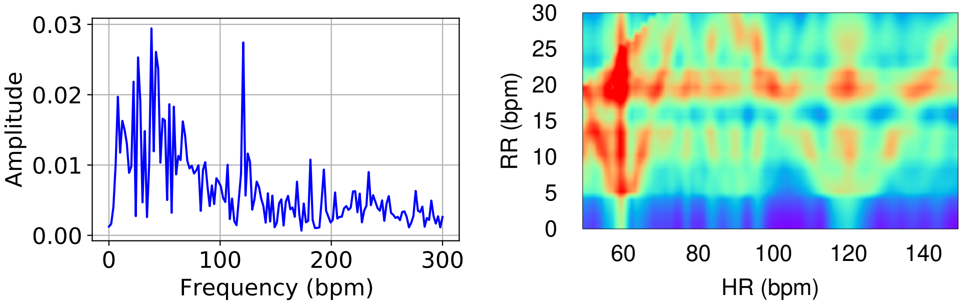
Next we describe how to detect whether the signal is corrupted beyond recognition, or vital signs are still "available". Large body motions (e.g., swaying) cause severe disruptions in the signal. Such "unavailable" signals must be detected and excluded to avoid producing erroneous results. Motion detection [5, 45, 90] based on periodicity in the time domain and/or condensed energy in the frequency domain have been proposed. However, strong respiration harmonics and intermodulation can dominate and mingle with such features from the much weaker heartbeat, and thresholding-based detectors cannot reliably tell them apart.

We propose a 2-D *"heatmap"* based detector that incorporates the spectral amplitudes at different frequencies. The heatmap $HM(f_r,\ f_h)$ borrows the concept of "joint probability distribution" and the value of each pixel is defined at the respiration/heart rate candidate pair $\{f_r,\ f_h\}$:

$$HM(f_r,\ f_h) = \sum_{z \in \mathbb{Z}(f_r,\ f_h)} A(z), \tag{12}$$

where $A(z)$ denotes the spectral amplitude of the signal at the frequency of z, and $\mathbb{Z}(f_r, f_h)$ is a set of potential harmonic and intermodulation frequencies, which can be expressed as $\{mf_h \pm nf_r | m, n \in \mathbb{N}_0\}$. When the signal is not corrupted much, harmonic and intermodulation frequencies of (f_r, f_h) close to the true respiratory and heart rates would have significant energy. Thus $HM(f_r, f_h)$ would gain relative large values of $A(z)$. This will visually appear as vertical and horizontal lines of large HM values in the heatmap. We show three representative samples in Figure 3, 4, and 5 for "available", partially "available", and "unavailable" signals. In Figure 3, the ground

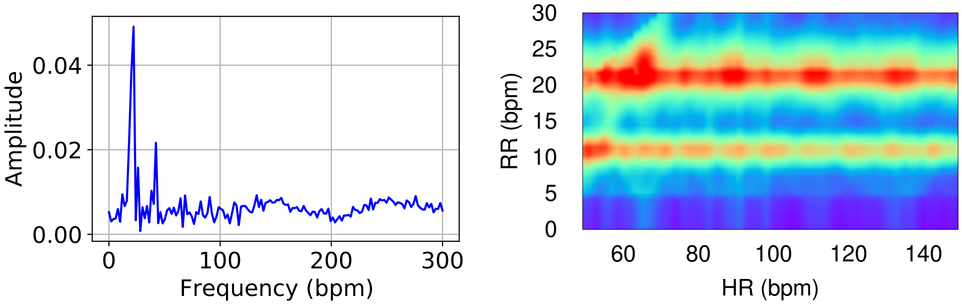
[3] The 5cm size is decided based on the amplitude of motion, the penetration effects of signals and errors in distance measurement.



(a) FFT Spectrum.

(b) 2-D Heatmap.

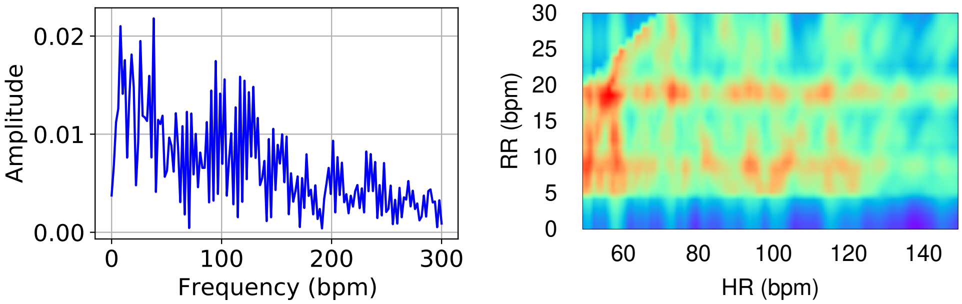
Fig. 3. An "available" sample: the heatmap shows an obvious red horizontal line near 20 bpm on RR, and a red vertical line near 60 bpm on HR. It matches the ground truth.



(a) FFT Spectrum.

(b) 2-D Heatmap.

Fig. 4. A partially "available" sample: the heatmap shows a red horizontal line near the 20 bpm ground truth RR but no strong vertical line around 75 bpm ground truth HR.



(a) FFT Spectrum.

(b) 2-D Heatmap.

Fig. 5. An "unavailable" sample: the heatmap is noisy and shows no obvious lines around ground truth RR and HR of 17 and 85 bpm.

truth RR and HR are 21 and 60 bpm. The heatmap shows a horizontal line near 20 bpm on RR and a vertical line near 60 bpm on HR with red color (i.e., larger values). Such visual patterns are used to detect whether a signal is "available". [4]

To learn the spatial-invariant features from the 2-D heatmap, we adopt the ResNet-18 model as the detector. The ResNet [28] model was initially proposed for image recognition, and takes 3-channel image data (i.e., RGB images) as input. We modify the first convolutional layer to process the heatmap, which is in the format of 1-channel grey-valued image. We also adjust the final layer (i.e., fc layer with softmax) to output a vector of two numbers (α, β), both within $[0,\ 1]$, indicating the normalized probabilities of availability and unavailability. The larger one determines the binary classification result of signal availability. Therefore, the probability of availability α can be used to indicate the signal quality.

The method of training and validation of the signal quality detector is described in §7.1. Signals detected as "available" are passed for vital signs estimation.

However, signals from the range bin that was directly located by the depth camera may not be suitable for vital signs estimation due to the offset error of the depth measure and the imperfect placement between UWB and depth sensors. Based on our preliminary experiments as described in §7.1 and especially in Figure 10, we note that adjacent range bins need to be considered to measure vital signs with better signal quality. To be specific, we flag a period as "available" when at least one range bin among 7 adjacent range bins (i.e., within ±15 *cm* range) is classified as "available". With α as the signal quality indicator, we select a range bin with the largest α among adjacent range bins for vital signs estimation during "available" period.

4.3 Vital Signs Estimation

As the phase of the UWB signal reflected from the chest wall changes corresponding to the physiological motions, we are able to extract the respiration and heartbeat. Given the center frequency of UWB signal 8.75 *GHz* in our design, a 0.2-0.5 *mm* displacement caused by heartbeat [68] translates to 2.1°–5.3° change in phase, while 4-12 *mm* displacement caused by respiration [19] translates to 42.0°–126.0° change in phase [38]. The heartbeat signal is orders of magnitude weaker than and totally buried by respiration signal in time domain. The difference in the typical frequency ranges between respiration (~6–18 *bpm*) and heartbeat (~50–150 *bpm*) allows them to be extracted separately. Figure 6 shows that with fine-tuned bandpass filters applied upon the phase signal, the respiration and heartbeat can be easily recognized from the FFT spectrum. While the example case looks straightforward, to robustly measure vital signs is still an open challenge. We will introduce estimation methods for respiration rate and heart rate respectively.

4.3.1 Respiration Rate Estimation. As the respiration frequency is usually from 0.1 to 0.3 *Hz*, we use a 2-order butterworth bandpass filter with a pass band of 0.1–0.8 *Hz* to remove the DC component and high frequency noise. Since the whole chest moves upon respiration, it has larger radar cross section (RCS) and displacement. Thus the phase signals are stable enough that we can easily estimate the respiration rate by counting the peaks. We use a time window of 30 *s* (which usually contains 5–8 breathing cycles), and calculate the time intervals between adjacent peaks. Then we average the interval to obtain the respiration rate f_r.

4.3.2 Heart Rate Estimation. Extracting the heart rate is more challenging due to its much smaller RCS and displacements, thus much weaker magnitudes in both temporal and spectral domains. As explained earlier (in §2.4), harmonics and intermodulation from respiration can easily dominate the heartbeat signal and their patterns are dynamic.

To robustly measure the heart rate, we propose a probabilistic weighted framework (PWF) that 1) incorporates four heart rate estimators each suitable to one of four identified temporal and spectral patterns; 2) adaptively

[4] Horizontal lines near 10 bpm on RR exist because the true 20 bpm respiration peak could be interpreted as second order harmonic. Still, such incorrect lines have weaker supporting evidences, thus smaller values and fainter colors.

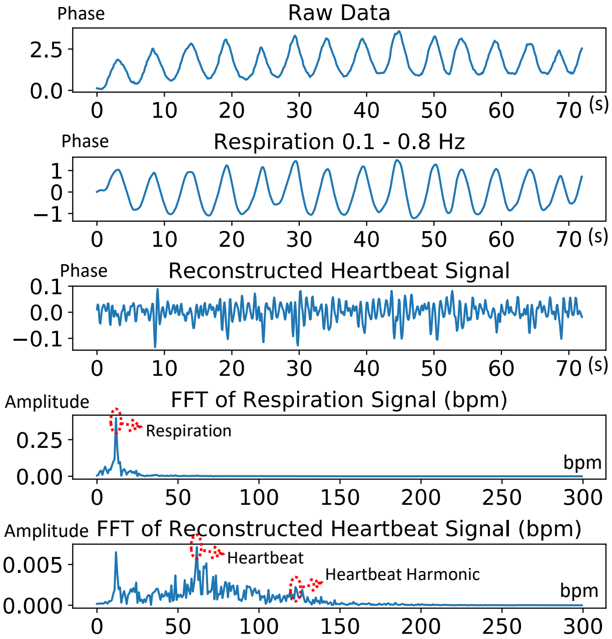


Fig. 6. An example segment of vital signal for spectral analysis.

combines heart rate candidates generated by the estimators with the quantified cumulative evidence of each pattern; and 3) leverages limits in heart rate temporal changes to smooth continuous measures.

Heartbeat Signal Extraction. In this step, we filter noises and respiration signals, and enhance the heartbeat signal for estimation. While the heartbeat signal presents periodic changes, the noise behaves randomly and can be modeled as Gaussian. We use auto-correlation [69] to zero-out the noise and enhance the periodic pattern of heartbeat. We observe that because of its higher frequency, heartbeat causes larger changes among adjacent sampling points than respiration. We use the second order difference to make the heartbeat more prominent.

Then we use the Discrete Wavelet Transform (DWT) as the filter bank [79] to extract heartbeat signals because DWT can retain the inherently irregular shape of the vital signals while the conventional filters (e.g, Butterworth filter [34]) would smooth the shape and result in loss of information for temporal analysis. We progressively split the signal into *approximation coefficients* (from the low-pass filter) and *detail coefficients* (from the high-pass filter) with the previously decomposed coefficients and reconstruct the signal with the coefficients in the interested frequency range (0.625–5 Hz, which covers both fundamental and second order harmonics). With L iterations (corresponding to L scales), an approximation coefficient $\gamma^{(L)}$ and a sequence of detail coefficients $\upsilon^{(1)}, \upsilon^{(2)}, ..., \upsilon^{(L)}$ are calculated in (13).

$$\begin{cases} \gamma_k^{(L)} & = \sum_{n \in \mathbb{Z}} s[n] \varphi^{(L)}_{2^L n-k}, L \in \mathbb{Z}, \\ \upsilon_k^{(l)} & = \sum_{n \in \mathbb{Z}} s[n] \psi^{l}_{2^l n-k}, l \in \{1, \dots, L\}, \end{cases} \tag{13}$$

where φ denotes the scaling function and ψ the wavelet. The heartbeat signal can be reconstructed using inverse DWT:

$$s[n] = \sum_{k \in \mathbb{Z}} \gamma_k^{(L)} \varphi^{(L)}_{2^L n-k} + \sum_{l=1}^{L} \sum_{k \in \mathbb{Z}} \upsilon_k^{(l)} \psi^{l}_{2^l n-k}. \tag{14}$$

In VitalHub, we select Daubechies(db4) wavelet as the mother wavelet [22], and split signal into 4 levels. The detail coefficients $\upsilon^{(3)} + \upsilon^{(4)}$ (ranging from 0.625 Hz to 2.5 Hz) are used to reconstruct the heartbeat signal.

The coefficients $v^{(4)} + \gamma^{(4)}$ (ranging from 1.25 Hz to 5 Hz) are used to reconstruct the second order harmonic component of the heartbeat signal.

Ensemble of Heart Rate Estimators. Based on manual examination of over 6000 data samples, we identify four typical temporal/spectral patterns (present in 98.55% of the data) and identify a suitable estimator of the fundamental heart rate based on each domain pattern, including 1) zero-crossing (ZC), 2) peak interval (PK), 3) local maximum detection in the spectrum of the heart rate range (LMD), and 4) spectral peak detection in the range of the heartbeat signals' second order harmonic (SOH).

The first two handles two temporal patterns. ZC estimates the heart rate by counting the number of zero-crossings in a time window, dealing with a periodic pattern of temporal changes between negative and positive values. Higher order harmonics of respiration may cause more negative-to-positive transitions, thus falsely higher heart rate. PK measures the average interval between adjacent local maxima in a time window, thus the heart rate. It is relatively immune to signals of larger energy, but sensitive to high frequency jitters.

The latter two handles two spectral patterns. When the fundamental spectral peak of heartbeat has significant energy [5], LMD detects such high peaks in the heart rate range (50−150 bpm). When higher order harmonics or intermodulation of respiration has strong energy, they may overwhelm the heart peak in this range. SOH selects spectral peaks in the range of the second order harmonic of the heartbeat (100-300 bpm), then halve them as estimates. We observe that respiration harmonics and intermodulation have much weaker energy in this range [64]. Due to partial overlap with the heartbeat fundamental frequency range, sometimes respiration may still produce significant peaks thus erroneous heart rate estimation.

Using a sliding window, we produce a heart rate candidate set C_t at time t, including C_t^F, 2 estimates from ZC, PK and 3 largest peaks from LMD, and C_t^S, 3 largest peaks from SOH. Unless explicitly stated, a candidate c_t^m is chosen from the combined set $C_t = C_t^F \cup C_t^S$.

Probabilistic Heart Rate Tracking. We formulate the continuous heart rate estimation as tracking the "trend" of changes, with the state update equation as follows:

$$\hat{x}_t = x_{t-1} + \dot{x}_{t-1} \triangle t + \varepsilon_p, \tag{15}$$

where x_{t-1} is the state (i.e., heart rate) we have estimated at time $t-1$, $\hat{x}_t$ is the heart rate predicted at time t, $\triangle t$ is the estimation interval (set to 1 second in our configuration), and $\varepsilon_p \sim \mathcal{N}(0, \sigma_p^2)$ is the process noise. Because errors accumulate over time, the predictions must be calibrated using evidences from observations.

The four temporal/spectral patterns are present most of the time (> 98%), thus the heart rate candidate set C_t very likely includes the correct one. The key is to determine which one. We quantify the evidence of each candidate c_t^m to determine its weight and calibrate predictions.

- *Respiration Harmonics.* Assume the fundamental respiration frequency is f_t^r, then its harmonics are represented as $H_t^r = \{f_t^r, 2f_t^r, ..., Nf_t^r\}$, where N is empirically limited at 5 because those beyond the 5th are negligible [64]. The closer a candidate is to any respiration harmonic, the less likely it is true, which can be formulated in the following weight:

$$P_r(c_t^m) = 1 - g_r(\min_n(abs(c_t^m - n \cdot f_t^r)), \tag{16}$$

  where $n \in \{1, 2, ..., N\}$, $g_r(\cdot) \sim N(0, \sigma_r^2)$ is a Gaussian distribution and σ_r is empirically set to 2.
- *Heartbeat Harmonics.* Heartbeat signal also has harmonics, while random noise and may not. Thus the existence of high order harmonics can be used as an evidence of the heart beat fundamental frequency f_h. As the heartbeat signal is relatively weak, we only consider its second order harmonic. This weight can be calculated as follows:

$$P_h(c_t^m) = g_h(\min_n(abs(c_t^m - c_t^n)), \tag{17}$$

$$P_h(c_t^n) = g_h(\min_m(abs(c_t^m - c_t^n)), \tag{18}$$

where $c_t^m \in C_t^F$, $c_t^n \in C_t^S$, $g_h(\cdot) \sim N(0, \sigma_h^2)$ is another Gaussian, and σ_h is empirically set to 2.

- *Peak Prominence.* We observe that real peaks are usually "sharp" (i.e., higher prominence), even though the amplitude may be small. While we have estimations from both the time domain (i.e., ZC, PK) and the frequency domain (i.e., LMD, SOH), we use the prominence of the spectral peaks of the heartbeat signal reconstructed according to (14) at the corresponding (estimated) frequencies to regulate their weights, because the spectral pattern (i.e., the distribution of the peak prominence) is resilient to noise and can serve as a reliable indicator for selecting the vital signs candidates estimated from either temporal or spectral methods. We use an exponential distribution to represent this weight:
$$P_p(c_t^m) = 1 - e^{-\alpha \cdot p(c_t^m)}, \tag{19}$$
where $p(c_t^m)$ is the peak prominence which quantifies how much the candidate c_t^m peak stands out due to its height and location relative to other nearby peaks, and the scale factor α is empirically set to 1.
- *Temporal locality.* The heart rate is not likely to change abruptly in a short time (e.g., one second), and the next heart rate is usually close to the current one. Therefore, we quantify how close a candidate is to previous estimation as:
$$P_l(c_t^m) = g(abs(c_t^m - x_{t-1})), \tag{20}$$
where $g_l(\cdot) \sim N(0, \sigma_l^2)$ is another Gaussian. σ_l is the variance of heart rate trend.

We define the likelihood of a candidate to be the heart rate as the cumulative evidence in a product form:
$$\mathcal{L}_t^m = P_r(c_t^m) \cdot P_h(c_t^m) \cdot P_p(c_t^m) \cdot P_l(c_t^m) \tag{21}$$
The normalized weight for a candidate is expressed as:
$$\omega_t^m = \frac{\mathcal{L}_t^m}{\sum_{j=1}^{M_t} \mathcal{L}_t^j}, m = 1, 2, ..., M_t, \tag{22}$$

Then, we take the weighted average of all the candidates as a new measurement:
$$\bar{c}_t = \sum_{c_t^n \in C_t} \omega_t^n \cdot c_t^n. \tag{23}$$
We observe that the error of the weighted measurement can be considered zero-mean Gaussian (using Kolmogorov-Smirnov statistic statistic found at 0.036, less than 0.05, the threshold when two distributions are considered the same [25]). Therefore, we apply Kalman Filter to iteratively repeat the following steps to update the heart rate at discrete time steps upon each new candidate set:
$$\begin{gathered} K_t = \frac{\sigma_{t-1}^2}{\sigma_M^2 + \sigma_{t-1}^2}, \\ \sigma_t^2 = (1 - K_t)\,\sigma_{t-1}^2, \\ x_t = \hat{x}_t + K_t(\bar{c}_t - \hat{x}_t) \end{gathered} \tag{24}$$
where K_t is the Kalman Gain, σ_M^2 and σ_t^2 are the variances of measurement noise (from $\bar{c}_t$) and process noise initialized with σ_p^2.

5 CONTEXT ANNOTATION

To enable context-aware cohabited monitoring, subject identities and activities must be labeled with the detected vital signs. The meaning of context annotation is two-fold for health analytics over continuous vital signs measurements: 1) Only when correctly associated with the corresponding identities, the recorded vital signs data can be used for meaningful customized analytics. 2) With the annotated activity context, it helps detect anomalous deviations from a user's normal distribution (e.g., a user's vital signs will rise during exercise, but will become stabilized during sleep), and reduce false alarms. We present an effective, privacy-friendly user identification approach based on human walking skeleton data, and a probabilistic model for continuous user identity tracking under occlusions. Leveraging the same set of features, we recognize the physical activities as context information for each individual.

5.1 User Identification

To capture both spatial and temporal features while a user is walking, we leverage the skeleton data from a short period (a few steps' walking when the user enters the monitoring zone) as input for recognition, rather than individual frames.

Features. We leverage the dynamic body joints locations tracked from depth sensor as features for user identification. We calculate the vectors between adjacent body joints to obtain the features $\mathbb{V} = \{\vec{v}_0, \vec{v}_1, ..., \vec{v}_{N-1}\}$, where $\vec{v}_i = [x_i - x_j, y_i - y_j, z_i - z_j]$ is the vector from joint i to joint j, and N is the number of pairs of joints we use. Accordingly, we can also calculate the lengths for all the vectors as $\mathbb{L} = \{l_0, l_1, ..., l_{N-1}\}$, where $l_i = |\vec{v}_i|$, and the angles at major joints whose angle changes indicate specific activities (e.g., certain changes in neck angle indicate nodding). $\mathbb{A} = \{a_0, a_1, ..., a_M\}$, where $a_i = cos^{-1}(\frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \cdot |\vec{v}_j|})$. We choose the combination of $\{\mathbb{V}, \mathbb{L}\}$ as features because it shows best performance in the evaluation. We choose a time window of 2 seconds (i.e., 120 frames with 60 fps) to balance the recognition latency and accuracy, and each training sample has a dimension of $(120 \times (3 + 1) \times N)$. To reduce noise in the feature data, we use a Savitzky-Golay filter [20] to smooth data and filter out jitters among adjacent frames.

Identification Model. We design a deep recurrent model with stacked-LSTM layers [30], each with 128 hidden units, and a fully connected layer with Softmax activation to output prediction results. We choose to use stacked-LSTM layers because it has been found to have better capacity to learn useful features from abstract sequential data [10]. We empirically tune and finalize our model with two stacked LSTM layers to balance the tradeoff between the model capacity and computing complexity based on our preliminary experiments. With batch normalization [65], our model is free of the vanishing gradient issue. We choose to use stacked-LSTM layers because it has been found to have better capacity to learn useful features from abstract sequential data [10]. We empirically tune and finalize our model with two stacked LSTM layers to balance the tradeoff between the model capacity and computing complexity based on our preliminary experiments. While the LSTM [30] itself is robust to the vanishing gradient issue by its nature, we apply batch normalization [65] in the training phase to further prevent and get rid of the vanishing gradient issue. We train the model using cross-entropy loss [18] and Adam optimizer [37]. With the input of a series of features set $\{\mathbb{V}, \mathbb{L}\}$, both static features (e.g., limb lengths) and dynamic features (e.g., the walking pattern) can be effectively used for reliable user identification.

5.2 Probabilistic Identity Tracking

The identities of each skeleton data must be tracked continuously. Naively running LSTM inference continuously is inefficient and error prone, and the situation becomes even worse when non-line-of-sight (NLOS) happens due to occlusions. We propose a lightweight probabilistic identity tracking algorithm which keeps tracking identities of each user from earlier LSTM inferences to reduce complexity and improve accuracy.

The basic intuition is that human movements are continuous, thus we can predict the trajectories based on previous locations and directions upon transient occlusions. Given 60 fps of depth sensor, the moving distances between two adjacent frames are small enough that the skeletons can be easily tracked. We only need to recover the identities when occlusions happen (e.g., one user blocks the line-of-sight of another).

As shown in Figure 8, when there's only one user A in the field-of-view (FoV), the identity can be tracked easily by tracking the skeleton. When both A and B are present, B may lose his identity if occluded by A when B is walking. The problem becomes more complex when more users are present. We formulate this into a probabilistic estimation problem leveraging the user movement trajectories: we predict the "next appear" locations of the

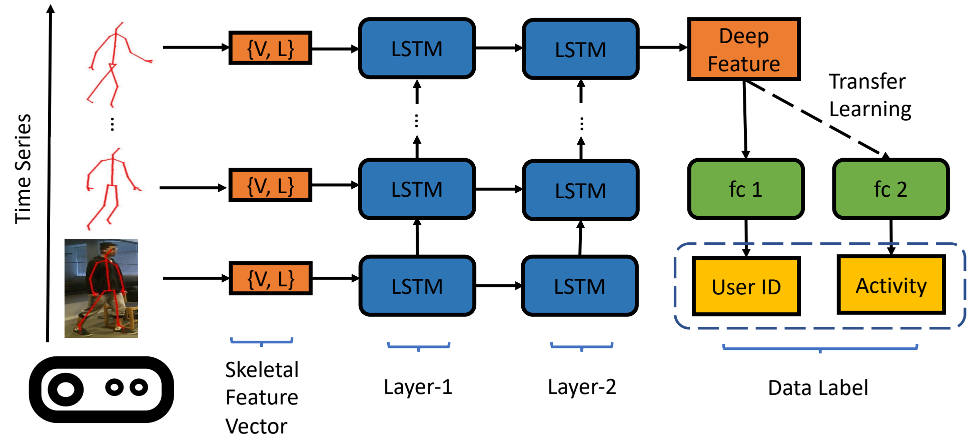


Fig. 7. Data labeling with skeleton data: the two stacked-LSTM layers take sequential data for feature extraction, followed by two parallel fully connected layers ("fc 1 & fc 2") for user identification and activity recognition.

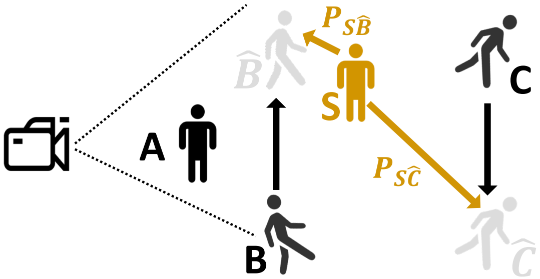


Fig. 8. An example of user identity tracking.

users when they are occluded, and estimate the identity probabilities based on how close the predicted “next appear” locations are to where the lost skeletons reappear in the FoV. Specifically, it works in three steps:

1. User Movement Prediction. We choose the user’s head location coordinate x_t^k to represent the k-th user’s location at time t. Given the user’s location x_{t-1}^k at time $t-1$ and the control signal $u_t^k = (v_t^k, \omega_t^k)$ where v_t^k is the moving speed and ω_t^k is the heading direction (estimated from the existing trajectory), the current predicted location $\hat{x}_t^k$ of k-th user is estimated as a displacement plus a Gaussian noise:

$$\hat{x}_t^k = x_{t-1}^k + u_t^k \triangle t + \varepsilon. \tag{25}$$

2. Identity Probability Update. When users are occluded, we keep predicting the “next appear” locations using Equation (25). Note that, the longer the user disappears, the more predictions we make, the lower the accuracy. As shown in Figure 8, user B and C walk in opposite directions, and are both occluded by user A. [5] $\hat{B}$ and $\hat{C}$ are the predicted locations for B and C, respectively. Suppose one skeleton S appears again at location x_t^S after the occlusion, we need to recover the identity for S. In such case, we denote the probability for S to be user B or C as $P_{S\hat{B}}$ and $P_{S\hat{C}}$, which can be estimated as follows:

$$P_{S\hat{B}} = g(\hat{x}_t^B) \cdot e^{-\alpha \cdot \tau}, \tag{26}$$

[5] Cases of more concurrent user occlusion are solved similarly and we do not repeat.

where $g(.) \sim N(x_t^S, \sigma^2)$ is a Gaussian distribution, $e^{-\alpha\tau}$ is a time decaying factor and τ is the eclipsed time of the occlusion. Similarly, we can estimate $P_{S\hat{C}}$.

3. User Identity Recovery. After we get $P_{S\hat{B}}$ and $P_{S\hat{C}}$, we first sort the probabilities and find the maximum one P_{max} and compare it to a threshold ϵ, which is the minimum probability we need to recover S. If $P_{max} < \epsilon$, it means we do not have enough confidence to recover S to either of the occluded users. In such cases, we need to run LSTM model to obtain the identity from skeleton features. In general, a larger ϵ makes identity recovery more robust, but is more compute expensive. If $P_{max} \geq \epsilon$, we further compare the difference between P_{max} and the second highest probability P_{second}. The user's identity is recovered only when $P_{max} - P_{second} \geq \eta$, where η is another threshold used to ensure sufficient difference in probabilities to avoid ambiguity among multiple identities. A larger η makes identity recovery more robust but incurs more computation to reach the threshold. We set $\epsilon = 0.8$ and $\eta = 0.5$ empirically to balance robustness and computation. In practice, once the users are recognized by initial LSTM inferences, the probabilistic tracking algorithm can track the identities efficiently, and invoking compute-heavy LSTM inferences only occasionally.

5.3 Activity Context Recognition

Knowing the concurrent activity (e.g., exercising or sleeping) is critical to detect health anomalies given the same vital sign changes (e.g., increased heart rates). We apply transfer learning with the LSTM model trained in §5.1 to recognize the activity. Since the pre-trained LSTM model has learned sophisticated features from sequential skeleton data, we feed them to a new fully connected layer (classifier) to recognize different activity categories. As shown in Figure 7, the sequential skeleton data feed the shared forward path of LSTM model, and the extracted feature vector generates predictions of both user identity and activity.

6 TESTBED

In this section, we describe the implementation of our testbed and experimental setup for evaluation.

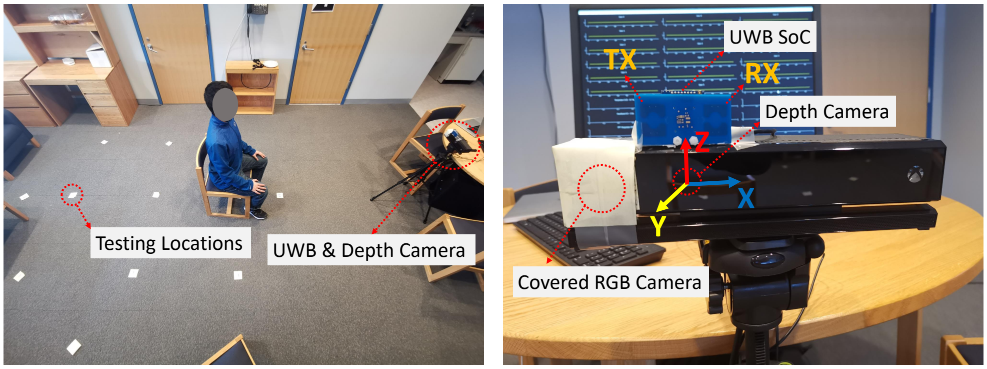
6.1 Implementation

VitalHub adopts a COTS IR-UWB sensor XeThru X4M03 [48] as its frontend for wireless sensing. The transmitted pulse is configured to be within tbe frequency band 7.25-10.2 GHz centered at 8.75 GHz, and the sampling frequency is 23.328 GHz. The frame rate of the UWB sensor is configured to be 10 frame-per-second (fps), and each frame includes samples of the echo pulses reflected from the objects within the range of 10 m. Kinect XBox serves as the depth sensor in VitalHub. Its SDK incorporates the human body pose recognition model [70] to detect human bodies present in the field of view at 60 fps. Both modalities stream data to the same backend PC via serial port. For computations of the whole pipeline, we use a laptop ROG Strix GL704 as the backend PC, which has an Intel i7-8750 2.2GHz CPU, 16GB RAM and NVIDIA RTX 2060 GPU. We implement deep learning models with PyTorch and run them using the GPU on the laptop.

6.2 Experimental Setup

Figure 9(b) shows the hardware setup of a Kinect XBox One sensor with *RGB camera covered* and a co-located UWB sensor. We conduct experiments in a room with a size of 4.5 × 9 m^2 (shown in Figure 9(a)).

We invited 8 students as participants for data collection (heights 156–192 cm, weights 49–108 kg), following a pre-established protocol that protected the anonymity of the students. We use two FDA approved medical devices, Nonin LifeSense II [55] and Masimo Pulse Oximeter [2] that can measure instantaneous heart and respiratory rates as the ground truth. We use the time stamps to obtain the alignment between such instantaneous vital signs estimation and the corresponding ground truth over time. Although results for each module are presented separately, VitalHub inherently integrates and produces data in a holistic pipeline concurrently.



(a) Testing environment setup.

(b) Hardware configuration.

Fig. 9. The experiment environment and hardware setup.

7 MICROBENCHMARKS

Before we delve into the end-to-end evaluation of VitalHub for vital signs monitoring, we start with a few microbenchmarks to demonstrate the performance of *signal quality detector* and *context annotation*. Then, to evaluate the end-to-end performance in §8, the data are retrieved according to the recognized identities based on *context annotation*; the pre-trained *signal quality detector* is used to filter in the time domain (i.e., sliding windows) and in the space domain (i.e., range bins) for robust vital signs estimation against inadvertent motions.

7.1 Signal Quality Detector

We first evaluate the signal quality detector for the classification of signal availability. Then we demonstrate how the detector can boost the performance of vital signs monitoring by reducing erroneous results from corrupted signals.

Classification. We compare the heatmap based detector (HM) against 4 existing detectors based on moving average (MABD) [90], moving variance (MVBD) [90], average variance energy (AVE) [45], and flat spectrum (FSD) [5].

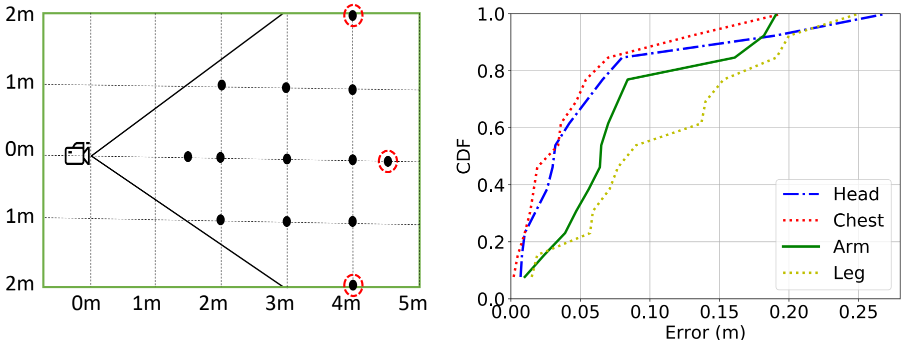
We build a balanced data set consisting of 20, 000 data samples, with equal number of "available" and "unavailable" samples randomly selected from 40, 782 manually labeled ones. Each data sample is the vital signals in a 30-second time window from one of 7 adjacent range bins centered at the depth sensor reported human body distance. We label a data sample as "available" if well-trained human observation identifies sufficient temporal periodicity and/or spectral peaks for both respiration and heartbeat, even under strong noises; otherwise, it is "unavailable". Therefore, for an identified "available" data sample, we know for sure the vital signs information exists. Thus failure to extract accurate readings indicates limitations of estimation algorithms.

We use precision (P), recall (R) and F-score ($= 2\frac{P\cdot R}{P+R}$) as metrics. Precision is the fraction of true positives among all identified positives, defined as $P = \frac{TP}{TP+FP}$; recall is the fraction of identified positives among all true positives, defined as $R = \frac{TP}{TP+FN}$. A high precision means unavailable data is unlikely to be falsely identified as "available"; and a high recall means the available data can be correctly identified thus utilized for monitoring. F-score quantifies the balance between precision and recall.

We apply 5-fold cross validation, and in each iteration we take 80% of the data set for training our detector or searching thresholds of others, and the rest 20% for testing. For fair comparison, each threshold is selected when respective F-score is maximized. The HM detector uses Adam optimizer [18] that minimizes cross entropy [37] as loss function, which measures the discrepancy between predicted and actual labels.

Table 1. Precision, recall and F-score of signal quality detectors.

	Precision (%)	Recall (%)	F-score (%)
MABD	93.11	47.80	63.17
MVBD	80.08	49.17	60.93
AVE	97.82	50.90	66.59
FSD	89.37	83.39	86.28
HM	96.41	96.29	96.35



(a) Room layout and measurement locations.

(b) CDF of error measurements.

Fig. 10. Kinect key body joints measurement accuracy.

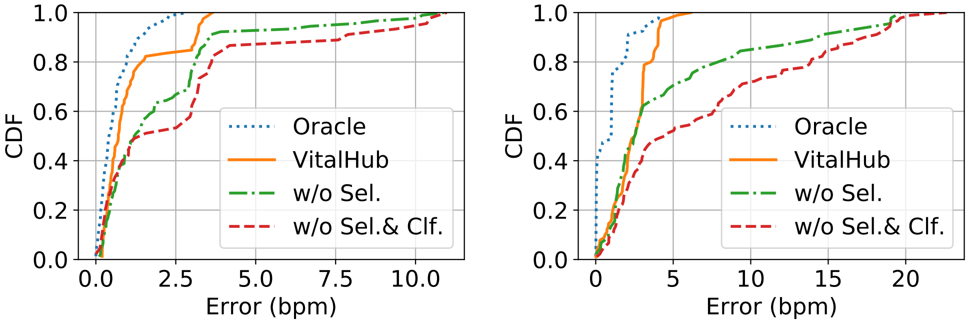
Table 1 shows that the time domain methods MABD, MVBD and AVE have relatively low recall. This is because the temporal signal is dominated by respiration signal in the shape, and sensitive to noises (e.g., environment, body movements). Achieving high precision requires "strict" selection, thus low recall and loss of available data. The frequency domain method FSD has better performance, but still about 10% worse than our HM detector. It assumes that the spectral peak sharpness (i.e., how condensed is the energy) indicates the availability of both respiration and heartbeat signals, however it is not always the case. Besides, respiration harmonics and intermodulation can also reduce the sharpness even if both respiration and heartbeat signals are available.

The HM detector requires more computation. The generation of and inference on the heatmap take 72.49 ± 6.99 *ms* and 7.26 ± 6.84 *ms* respectively, short enough for real-time measurements updated every 1 second.

Range Bin Selection. We observe that the distance reported by the depth sensor may not give the range bin with the best signal quality. To test the depth sensor's accuracy, we select a set of locations within the 4.5 *m* of the depth camera (Figure 10(a)). We ask the testing volunteer stand at each location, and we measure the ground truth distance to the head, chest, left arm and left leg using a laser measurement tool.

The errors of the depth sensor for four body parts are shown to be 10 *cm* at 80-percentile (Figure 10(b)), which is about twice the size of a range bin (5.14 *cm*). Thus we search 7 adjacent range bins (±15 *cm*) centered at the depth camera reported one, and select the one with the highest signal quality indicator α (provided by the trained HM detector).

Ablation Study. To study the effectiveness of the signal quality detector on the end-to-end system, we compare the performance in vital signs estimation with the progressive ablation of range bin selection (Sel.) and availability classification (Clf.) against an impractical "Oracle" that always knows whether the signal is available, which is the best range bin, and best estimator (among the four used in PWF) at each moment.



(a) CDF of RR error.

(b) CDF of HR error.

Fig. 11. Boosted performance with signal quality detector.

Figure 11 shows obvious performance degradation each time bin selection or classification is removed. With both of them, VitalHub achieves end-to-end respiratory and heart rate estimation at 1.5/3.2 bpm errors at 80-percentile, very close to 1.2/1.5 bpm errors by the idealistic oracle. This shows the necessity of the detector, which enables VitalHub to approach the "ceiling" of the oracle.

7.2 Context Annotation

We first compare the accuracy of different features and machine learning models (§7.2.1), then the effectiveness of identity tracking under occlusions (§7.2.2), finally the accuracy of activity recognition (§7.2.3).

7.2.1 User Identification. Considering VitalHub targets in-home deployment, we invite 8 volunteers (use as the maximum number of family members) to collect walking skeleton data for evaluation of user identification. Each contributes 5 *mins* data, resulting in a total of 18,000 frames at 60 fps. To generate the training data, we use a time window of 2 seconds with a step of 0.1 *s*. Thus the total training data size is around $\frac{5\times60s}{0.1s}\times 8 = 24000$ samples. Another 2 minutes' data from each volunteer are collected *separately* as testing data.

Precision, Recall, F-score. We adopt the same metrics as in §7.1 to evaluate the identification model. Table 2 shows the results of different feature combinations (described in §5) under our LSTM model. The combination $\{\mathbb{V},\mathbb{L}\}$ outperforms all others, thus selected as final features.

Table 2. Precision, recall and F-score of different features.

	Precision (%)	Recall (%)	F-score (%)
$\{\mathbb{V}\}$	85.60	85.12	85.01
$\{\mathbb{V},\mathbb{A}\}$	63.24	57.80	56.19
$\{\mathbb{V},\mathbb{L}\}$	86.98	85.49	85.37
$\{\mathbb{V},\mathbb{L},\mathbb{A}\}$	80.47	77.97	78.14

Different Classifier Models. We compare the performance of different classifiers using the same test data set with the feature of $\{\mathbb{V},\mathbb{L}\}$. Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayesian (NB) and Support Vector Machine (SVM) are used as baseline classifiers to compare with our LSTM based model. Notably, all the models have been carefully examined and configured with optimal settings regarding the respective parameters of interest. Specifically, LDA is configured with a threshold of 0.0002 for the solver of Singular Value Decomposition; KNN is configured with the distance metric of Minkowski and the number of neighbors of 5; DT is configured with the Gini impurity as the criterion of split; NB is configured with

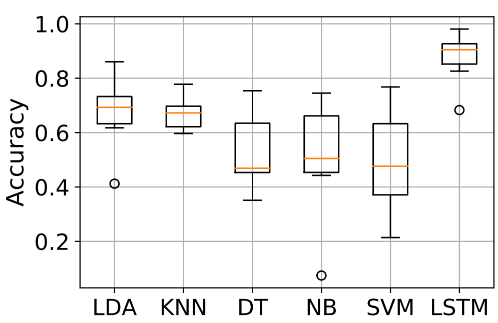


Fig. 12. Different classifiers.

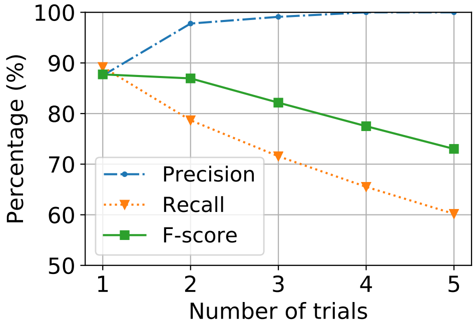


Fig. 13. Performance with different number of trials.

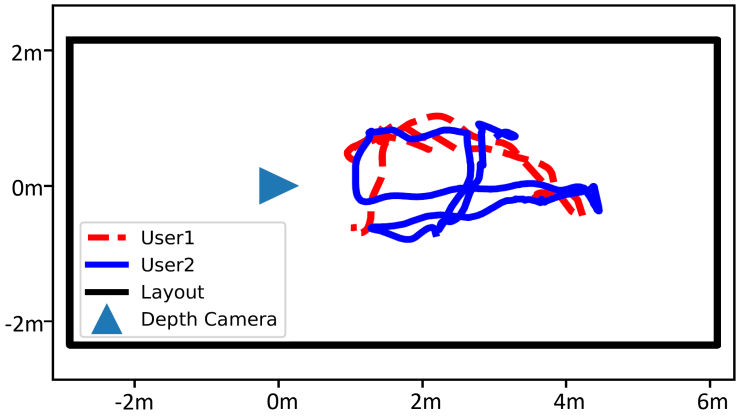


Fig. 14. Traces of identities being tracked in the presence of occlussions.

Gaussian distribution; SVM is configured with the regularization parameter C of 1.0 and tolerance for stopping of 0.001; LSTM is configured with the number of stacked layers of 2. Figure 12 shows the accuracy of each model. Our LSTM based model outperforms baseline significantly with a median accuracy around 90%. The only outlier in LSTM result is caused by excessive movements during data collection, causing large variations in raw data.

Different Number of Trials. To avoid "pollution" from false identifications, the model makes decisions only when a successive N trials produce the same result, at the cost of lower recall. Figure 13 shows the precision, recall and F-score with different number of trials. We choose 3 trials as the precision is close to 100% and the recall is acceptable to still generate enough data in longitudinal monitoring. In such setting, 3 LSTM prediction cycles take about 6 s to recognize a user.

7.2.2 Identity Tracking and Maintenance. To evaluate the robustness of our probabilistic identity tracking, we asked volunteers to simulate the occulsion cases discussed in §5.2. Results show our algorithm can resume the identities instantly when users reappear after temporary occlusions ($\sim$ 2 seconds). Figure 14 shows one example when two participants walk across the FOV of the depth camera, their respective traces are being tracked correctly and continuously even they occlude each other. The LSTM re-identification after long occlusions require multiple cycles thus longer time (e.g., 3 cycles take $\sim$ 6 seconds), and re-identification succeed for all cases we test. These show our identity tracking algorithm is effective and robust.

7.2.3 Activity Context Recognition. We collect 32,828 and 19,478 samples, with the same format as in the user identification, for training and testing data of the following activities: eating (ET), lying (LY), running (RN),

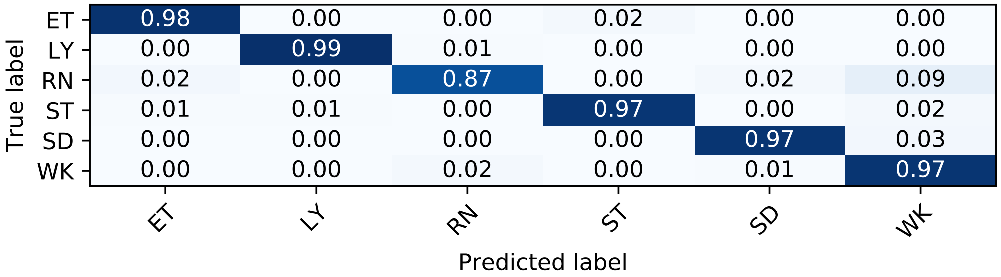
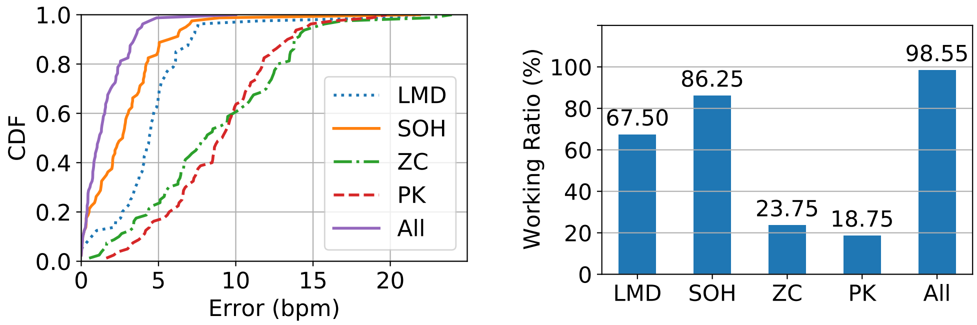


Fig. 15. Activity recognition confusion matrix.



(a) CDF of HR error.

(b) Working ratio of different HR estimators.

Fig. 16. Performance of individual estimators.

sitting (ST), standing (SD), and walking (WK). Every participant is asked to perform each category of activities repeatedly during data collection. The results are shown in Figure 15. Overall, it achieves 96.27% precision, 96.09% recall, and 96.10% F1-score. A small portion of running samples are falsely recognized as standing or walking for transitions in between. Such performance is sufficient for longitudinal in-home monitoring.

8 END-TO-END VITAL SIGNS MONITORING

In this section, we compare different methods in estimating vital signs and dealing with non-linearity issues (e.g., harnomics, intermodulation, and dynamic signal patterns as described in §2.4), then we study the impact of user and environment factors to end-to-end vital sign monitoring performance.

8.1 Estimators and Non-linearity Study

8.1.1 Individual Estimators. We evaluate the effectiveness of all the 4 estimators for heart rate estimation (Figure 16). We define a "working ratio" metric as the fraction of time when an estimator has < 5 *bpm* error, which is an acceptable error range for long-term monitoring. The second order harmonic estimator (SOH) has the highest working ratio, because the second order harmonic of the heart rate is spectrally free of the high order harmonics of respiration. Temporal methods zero-crossing (ZC) and peak interval (PK) have relatively low working ratio due to sensitivity to noise and interference. However, they produce more gradual changes in output compared to spectral methods LMD and SOH, helping avoid large jumps between spectral peaks for smooth tracking. VitalHub combines all of them and achieves over 98% working ratio. This demonstrates the effectiveness of the PWF framework combining less reliable estimators to achieve more robust estimation.

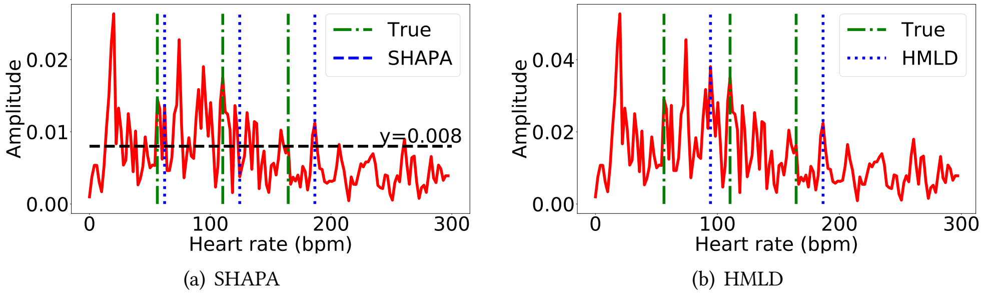


Fig. 17. Typical spectrum when SHAPA and HMLD both fail

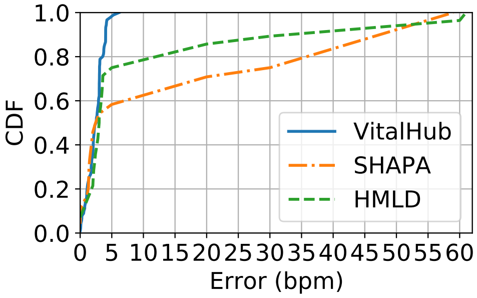


Fig. 18. VitalHub, SHAPA and HMLD on HR estimation.

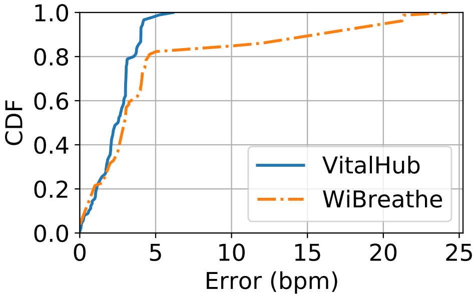
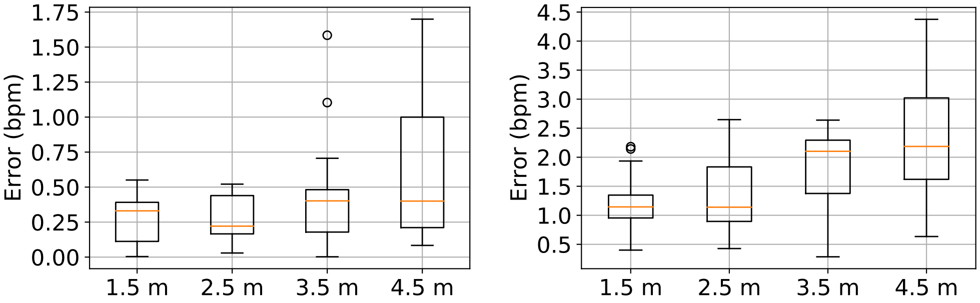


Fig. 19. VitalHub and Wibreathe on HR estimation.

8.1.2 Robustness against Harmonics and Intermodulation. We identify two representative methods, SHAPA [53] and HMLD [93], dealing with harmonics and intermodulation and compare with them. They leverage the frequency relation between harmonics. SHAPA tries to find three spectral peaks (i.e., "harmonic path") with 1 : 2 : 3 ratio in frequency with magnitude larger than some preset threshold. HMLD tries to find a pair of stable spectral peaks with 1 : 2 ratio in frequency. Figure 18 shows while VitalHub reaches 98% working ratio, SHAPA and HMLD deliver only 51.2% and 76.3% respectively. Both methods rely on presumed signal patterns, which may not always happen in reality.

SHAPA is very sensitive to SNR. The preset threshold is supposed to filter out most noise peaks while leaving those from fundamental and harmonics of heartbeat. However, when SNR is low, even with a well tuned threshold (one that just below all harmonic peaks), noise can easily cause incorrect estimation. Figure 17(a) shows the threshold set at the minimum magnitude of all harmonic peaks. However, many noise peaks exist above the threshold, and some present a better 1 : 2 : 3 relation than the real heart rate and its harmonics. Thus the algorithm chooses an incorrect harmonic path (62, 124, 186) instead of the true harmonic path (56, 110, 164). HMLD has similar problems: Figure 17(b) shows a wrong estimation (94) which is the intermodulation of heartbeat (56) and respiration (19). The above shows that designs relying on presumed signal patterns are not robust enough.

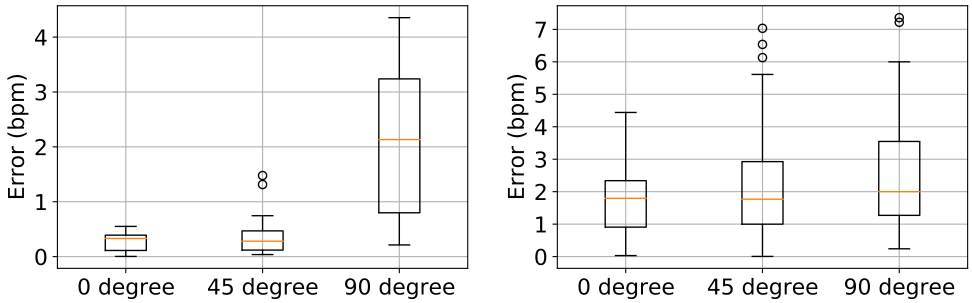
8.1.3 Dealing with Unpredictable and Dynamic Signal Patterns. We implement WiBreathe [61] for comparison as it is a most related work that identifies and addresses unpredictable and dynamic vital signal patterns. We caution that WiBreathe was designed for respiration only, so the comparison serves not to criticize, but to shed light on how applicable its techniques are for heart rate. WiBreathe adaptively combines several estimators' output, under the assumption that the majority of them would produce correct estimations. For fairness, we compare only the strategies in combining estimator outputs, while all other components such as preprocessing pipelines are the same. Figure 19 shows that the working ratio of WiBreathe can be up to 81.3% of the time, much lower than VitalHub's 98%. We find the majority of the HR candidates from the estimators can be incorrect, causing



(a) Impact of distances on RR estimation.

(b) Impact of orientations on HR estimation.

Fig. 20. Vital signs estimation under different distances.



(a) Impact of orientations on RR estimation.

(b) Impact of orientations on HR estimation.

Fig. 21. Vital signs estimation under different orientations.

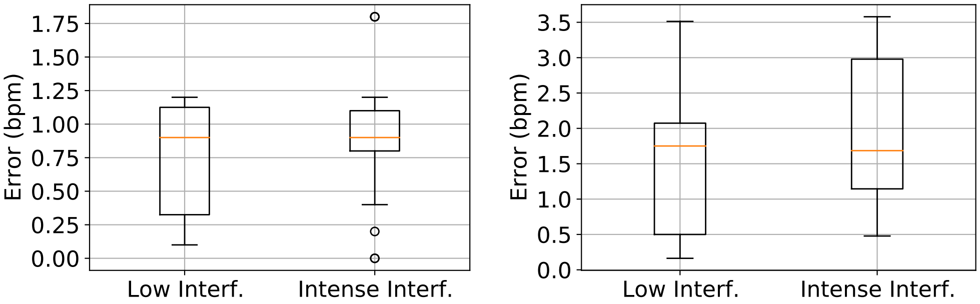
WiBreathe fail to make the correct estimation. Our PWF strategy uses the cumulative evidence thus can still select the correct candidate even it is not in the majority but possessing stronger evidence, thus dealing with dynamical signal patterns more effectively.

8.2 User and Environment Factors

8.2.1 Impact of Distances. We vary the distance between 1.5–4.5 *m* with a step length of 1 *m*, while keeping the orientation of the subject at 0 degree (facing frontally). The results are shown in Figure 20. We can see that respiratory and heart rate estimations are very stable even at 4.5 *m*, up to the range of the depth sensor.

8.2.2 Impact of Orientations. We vary the orientation of the subject at 0, 45, and 90 *degrees* while keeping 2.5 *m* distance (shown in Figure 21). Interestingly, we observe that HR accuracy is not affected much by the orientation, but RR error at 90 *degree* more than triples. This is because the breathing chest movement in the mediolateral axis (i.e., side) is only around 0.6–1.1 *mm* [77], much smaller than that in the front, thus more susceptible to errors.

8.2.3 Impact of Ambient RF sources. To evaluate the impact of ambient RF sources, we compare the performance in two settings: low Wi-Fi traffic where Wi-Fi signal comes from nearby buildings but no Wi-Fi device running indoors, and intense Wi-Fi traffic where 3 Raspberry Pis, 4 laptops, 4 smartphones and 2 Wi-Fi routers keep streaming data indoors. Figure 22 shows negligible decrease in the respiration and heart rate measurement. This



(a) Impact of ambient RF sources on RR estimation.

(b) Impact of ambient RF sources on HR estimation.

Fig. 22. Impact of different ambient RF environment.

is because UWB spreads the energy on a wide frequency bandwidth thus narrow band Wi-Fi signals in 2.4/5 *GHz* do not present severe interferences.

8.2.4 Multi-User Minimum Resolvable Distance. The minimum resolvable distance, i.e., how close adjacent subjects can be without interfering each other, is a very critical factor for co-habiting scenarios. We invite 3 volunteers, separated at 1 *m* initially, gradually decreased at 10 *cm* steps, until any two of them appear identical in measurement, meaning they are too close for the system to differentiate. We find all 3 volunteers can be reliably monitored even when separated at only 20 *cm*, with performance comparable to the single user setting.

9 DISCUSSION

Extensible Framework. In our vital sign monitoring module, we combine several estimators' output and leverage prior knowledge about vital signs to produce a weighted sum estimation to deal with the challenges from harmonics and intermodulation. Our framework can easily accommodate more estimation methods and other prior knowledge to improve the performance as advances are made in these fields. Evidence for signal patterns suitable for such methods will be quantified to update respective weights.

Body Motions. VitalHub measures vital signs only in quasi-stationary settings (e.g., talking, typing, watching TV) where the displacement of the chest wall due to physiological movements (i.e., heartbeat and respiration) are the main sources of distance thus phase changes. However, in non-stationary settings, large body movements (e.g., swaying of body) can overwhelm physiological movements, causing severe disturbances to phase changes and making it impossible to directly extract vital signs. We adopt a signal quality detector to flag the periods where the signal is "unavailable" to exclude corrupted signals that can degrade the accuracy of vital sign measurements. In longitudinal monitoring, even after discarding such corrupted data, there is still sufficient amount to detect health status changes because of the continuous, long term measurements.

Recently regular RGB cameras or a pair of radio sensors before and behind the body are used to measure and cancel random body movements [27, 44]. Due to privacy reasons, in our surveys people have expressed serious concerns on regular cameras. Using a pair of radio sensors may constrain the space where users can stay, but the general approach combining signals from multiple sensors seems promising dealing with large movements. We plan to investigate how to allow unconstrained movements in-home while dealing with large motions.

Trade-offs between Precision and Recall. We observe that for small fractions of time (less than 4%), the signal quality detector may fail. When this happens, none of the estimators can produce a correct heart rate estimate. If this continues long enough, the PWF may fail to smooth out such erroneous output and converge to some wrong heart rate estimate. This problem can be alleviated by combining signals in consecutive time windows, but at the cost of reduced recall of the data. We will study this in the future to see how to achieve a proper balance.

Sensing Range and Orientation. The effective range of the current prototype is limited by the UWB and depth sensors we use. The depth sensor has a LOS distance of 4.5 meters, whereas the UWB one has a range of near 10 meters (with some penetration capability). The range is limited by the lesser of the two. Nevertheless, it is still sufficient for room-size monitoring in-home with the system mounted on the wall thus most of the area covered in LOS. Our identity tracking algorithm can also help mitigate blockage due to cross-walking. We will explore depth sensors of longer ranges and radio-only solutions for context annotation.

The current prototype also produces higher errors in respiration rate when the orientation of the subject is near 90 degrees. This is mainly because the chest movement in the mediolateral dimension (0.6-1.1 *mm*) is comparable to the displacement attributed to the heartbeat [77]. We will study multi-sensor deployment so the 90-degree orientation of the subject can be avoided by at least one sensor.

Azimuth and Elevation Separation. The current prototype can only separate signals of different distances. Two people at same distance but different orientations are not separated. This is because the UWB sensor we use has only a single pair of transmitter and receiver, so it cannot differentiate azimuth and elevation. Proper placement of the sensor kit is needed to reduce the chances when subjects are at similar distances. Some research [91] has studied how to extract vital signs from entangled signals with blind signal separation [17] (e.g., Independent Component Analysis [16]). We will explore the robustness of such approaches, and MIMO, multi-sensor configurations for azimuth and elevation separation.

More Subjects and Real Patients. The diversity of our 8-subject pool is a bit limited. We deployed VitalHub in our university hospital for pilot test. However, due to COVID-19 we are unable to obtain sufficient data. We will resume the study on larger subject pools and real patients once the pandemic eases enough.

10 RELATED WORK

We describe the main work in vital signs monitoring and context information recognition, and how VitalHub compares.

Vital Signs Monitoring. Traditional medical equipment such as electrocardiography (ECG) and echocardiography [26] are accurate but too expensive and difficult to operate in home. VitalMon [35] uses geophones to sense bed vibrations from ballistic forces of sleeping subjects. Wearables using photoplethysmogram (PPG) sensors (e.g., pulse oximeters [2], watches [1] and wrist bands [23, 33]) require continuous skin contact and regular recharging. Some inertial sensor based methods [6, 49] place smartphones on the chest. Such touch-based sensing restrains the free roaming of users [32], presents cognitive and physical challenges for older adults living alone, thus not suitable for longitudinal passive in-home monitoring.

Researcher have proposed four types of methods for contactless ubiquitous vital sign monitoring: remote PPG [32, 62], acoustic [51, 60, 75, 87], WiFi [45, 46, 58, 79] and other RF based methods [15, 52, 72, 89]. Remote PPG measures the heart rate from an RGB video of the user's face [32]. It has privacy issues and may be impacted by skin color, make-up and lighting. Both passive [75, 87] and active [51] acoustic methods leveraging smart speakers or phones have been studied for respiration rate monitoring. Acousticcardiogram (ACG) [60] uses frequency modulated continuous wave (FMCW) sonar front end smartphones to monitor both heart and respiratory rates. It targets near-field monitoring at a maximum distance of 30 cm.

WiFi based methods leverage channel state information (CSI) [45, 46, 79] or received signal strength (RSS) [3, 31, 58]. RF methods have exploited techniques including mmWave [15, 89], doppler radar [52, 72], FMCW radar [5, 82], and UWB radar [43, 71, 84, 88]. Some of them monitor only respiration [3, 31, 52, 58, 72]. We observe respiration harmonics and intermodulation disrupt spectral based heart rate estimation severely, yet we do not see this described or treated sufficiently in the above work. We use a UWB sensor which has extremely short pulses and wide bandwidth, making it inherently more robust to interference and multipath [40].

The challenge of harmonics and intermodulation is analyzed in [39]. Work in electronics engineering community [21, 36, 53, 64, 93] has proposed methods based on certain assumptions of the signal's temporal, spectral patterns (e.g., magnitudes between fundamental and harmonic components, gradual changes in the heart rate). We observe such patterns are far from stable, thus these methods (e.g., finding pairs of stable spectral peaks with 1 : 2 ratio in frequency) often fail. WiBreathe [61] adaptively selects an output from multiple respiration estimators closest to the previous estimate. It assumes at least one estimator gives good estimation, which we find does not hold for heartbeat due to much weaker spectral energy, thus easily dominated by respiration harmonics and intermodulation. We combine multiple estimators by quantifying respective evidence of their suitable patterns in a probabilistic framework to enable robust heart rate tracking.

Most radio sensing based work targets quasi-stationary settings, while non-stationary settings remain an open issue [83, 86]. To detect large motions, many methods use a fixed threshold for phase change or spectrum sharpness [5, 29, 45, 51, 87, 90]. Motion-related sensors [57] and RGBD-cameras [54] are used as well. Such additional sensors add to the system cost and complexity, and fixed thresholds cannot handle complex signal dynamics. Our heatmap feature incorporates the full spectral characteristics of the vital signs and uses a deep neural network to achieve near human performance.

There are efforts to directly measure and cancel motion disruptions using accelerometers [9, 59], regular RGB cameras and radio sensor pairs before/behind the body [27, 44]. They require wearable sensors, or cause privacy concerns and constraints on the user's free roaming, thus not preferred for long-term in-home monitoring. We plan to explore multi-sensor collaboration to keep robust measurements under motion while retaining passive, low-cost sensing.

Context Information Recognition. Body pose thus activity recognition by non-touch means has been an active research area. Lots of work leverages visual data, including RGB images, videos, and depth [12, 47]. OpenPose [12] tracks body poses with a single RGB image or reconstructs 3D skeletal poses with multiple cameras. FaceNet [66] extracts facial features [97, 98] from RGB images for identification. Caesar [47] detects complex activities with multiple non-overlapping cameras. Despite the technology maturity, many people are uncomfortable living under cameras due to privacy concerns.

Inertial sensors have been used for tracking activities including dancing, smoking and exercise [13, 24, 56]. Visible light is used for skeleton reconstruction [41] with instrumentation of photodiodes or LED panels on the ceiling or floor, which may not be convenient in-home.

WiFi [73, 74, 78, 80, 81, 92, 96] and other RF [4, 14, 42, 94] are studied for human activity and motion sensing. E-eyes [80], mD-Track [81] uses WiFi signals to recognize and track indoor human activities. CrossSense [92] applies transfer learning to effectively reuse the learned knowledge across different sites. CARM [78] correlates WiFi CSI dynamics and human activities for recognition. Widar3.0 [96] estimates velocity profiles of gestures for cross-domain gesture recognition. WiMU [73] recognizes the gestures of multiple users simultaneously, and WiAG [74] recognizes the gestures irrespective of the user's position or orientation. Among RF-based methods, EAR [14] uses ambient RF signals for human activity sensing. Li et al. [42] feed RFID data into a CNN for activity recognition. RF-Pose3D [94] and RF-Capture [4] track the 3D positions of a person's skeleton even under full occlusion from the RF sensor.

We choose depth camera in the current prototype because it can produce mature and robust skeletal pose tracking [67], and easily acceptable by people due to lack of fine-grained visual features. It offers distance report, skeleton tracking, user identification and activity recognition simultaneously. RF-Pose3D [94] and RF-Capture [4] achieve such goals using customized FMCW radios with antenna arrays. We plan to study how to extend our UWB based system for robust and low-cost context recognition.

11 CONCLUSION

We present VitalHub, a robust, non-touch, passive sensing system for longitudinal in-home vital signs monitoring leveraging UWB and depth sensors. We describe how respiration harmonics and intermodulation cause strong disturbances to robust heart rate monitoring. We propose a probabilistic weighted framework that adaptively combines an ensemble of estimators based on the quantified cumulative evidence of their suitable temporal and spectral signal patterns. Besides, we introduce a LSTM-based neural network and a probabilistic tracking model to provide reliable context annotation (i.e., identification and activity recognition) using privacy-preserving skeleton features. Extensive experiments show that *VitalHub* achieves 1.5/3.2 "breaths/beats per minute" (denoted by "bpm") errors at 80-percentile for RR/HR, approaching the 1.2/1.5 bpm error "ceiling" of an idealistic but impractical oracle. We also share insights on why existing methods do not handle harmonics and intermodulation well. In addition, our context annotation module achieves 90% median accuracy for differentiating 8 subjects (while the number of cohabiting persons in one home would usually be less than 8) based on skeletal walking patterns, and above 96% precision for classifying among 6 common daily activities. With the automatic context annotation, the vital signs record can be valuable for future customized analytics (e.g., the detection of anomalous changes in vital signs from a user's normal distribution of daily routines/activities). We believe VitalHub offers a suitable solution for longitudinal in-home vital signs monitoring.

REFERENCES

[1] [n. d.]. Apple Watch. https://www.apple.com/watch/

[2] [n. d.]. Masimo - MightySat Rx. https://www.masimo.com/products/monitors/spot-check/mightysatrx/

[3] Heba Abdelnasser, Khaled A Harras, and Moustafa Youssef. 2015. UbiBreathe: A ubiquitous non-invasive WiFi-based breathing estimator. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 277–286.

[4] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[5] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 837–846.

[6] Heba Aly and Moustafa Youssef. 2016. Zephyr: Ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[7] Nikolaj Andersen, Kristian Granhaug, Jørgen Andreas Michaelsen, Sumit Bagga, Håkon A Hjortland, Mats Risopatron Knutsen, Tor Sverre Lande, and Dag T Wisland. 2017. A 118-mW pulse-based radar SoC in 55-nm CMOS for non-contact human vital signs detection. *IEEE Journal of Solid-State Circuits* 52, 12 (2017), 3421–3433.

[8] Etienne Antide, Mykhailo Zarudniev, Olivier Michel, and Michael Pelissier. 2020. Comparative Study of Radar Architectures for Human Vital Signs Measurement. In *2020 IEEE Radar Conference (RadarConf20)*. IEEE, 1–6.

[9] H Harry Asada, Hong-Hui Jiang, and Peter Gibbs. 2004. Active noise cancellation using MEMS accelerometers for motion-tolerant wearable bio-sensors. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 1. IEEE, 2157–2160.

[10] Rana Azzam, Yusra Alkendi, Tarek Taha, Shoudong Huang, and Yahya Zweiri. 2020. A stacked LSTM-based approach for reducing semantic pose estimation error. *IEEE Transactions on Instrumentation and Measurement* 70 (2020), 1–14.

[11] Lynn Bickley and Peter G Szilagyi. 2012. *Bates' guide to physical examination and history-taking*. Lippincott Williams & Wilkins.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[13] Keng-Hao Chang, Mike Y Chen, and John Canny. 2007. Tracking free-weight exercises. In *International Conference on Ubiquitous Computing*. Springer, 19–37.

[14] Zicheng Chi, Yao Yao, Tiantian Xie, Xin Liu, Zhichuan Huang, Wei Wang, and Ting Zhu. 2018. EAR: Exploiting uncontrollable ambient RF signals in heterogeneous networks for gesture recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 237–249.

[15] Huey-Ru Chuang, Hsin-Chih Kuo, Fu-Ling Lin, Tzuen-Hsi Huang, Chi-Shin Kuo, and Ya-Wen Ou. 2012. 60-GHz millimeter-wave life detection system (MLDS) for noncontact human vital-signal monitoring. *IEEE Sensors Journal* 12, 3 (2012), 602–609.

[16] Pierre Comon. 1994. Independent component analysis, a new concept? *Signal processing* 36, 3 (1994), 287–314.

[17] Pierre Comon and Christian Jutten. 2010. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

[18] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134, 1 (2005), 19–67.

[19] Anne De Groote, Muriel Wantier, Guy Chéron, Marc Estenne, and Manuel Paiva. 1997. Chest wall motion during tidal breathing. *Journal of Applied Physiology* 83, 5 (1997), 1531–1537.

[20] Yong Du, Wei Wang, and Liang Wang. 2016. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 95–108.

[21] Raghed El-Bardan, Dhaval Malaviya, and Albert Di Rienzo. 2017. On the estimation of respiration and heart rates via an IR-UWB radar: An algorithmic perspective. In *2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*. IEEE, 1–5.

[22] Ergun Ercelebi. 2004. Electrocardiogram signals de-noising using lifting-based discrete wavelet transform. *Computers in Biology and Medicine* 34, 6 (2004), 479–493.

[23] Biyi Fang, Nicholas D Lane, Mi Zhang, Aidan Boran, and Fahim Kawsar. 2016. BodyScan: Enabling radio-based sensing on wearable devices for contactless activity and vital sign monitoring. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 97–110.

[24] Abu Zaher Md Faridee, Sreenivasan Ramasamy Ramamurthy, HM Hossain, and Nirmalya Roy. 2018. HappyFeet: Recognizing and assessing dance on the floor. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*. ACM, 49–54.

[25] Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* 10, 2 (2012), 486.

[26] Nira A Goldstein, Nancy Sculerati, Joyce A Walsleben, Nasima Bhatia, Deborah M Friedman, and David M Rapoport. 1994. Clinical diagnosis of pediatric obstructive sleep apnea validated by polysomnography. *Otolaryngology?Head and Neck Surgery* 111, 5 (1994), 611–617.

[27] Changzhan Gu, Guochao Wang, Yiran Li, Takao Inoue, and Changzhi Li. 2013. A hybrid radar-camera sensing system with phase compensation for random body movement cancellation in Doppler vital sign detection. *IEEE Transactions on Microwave Theory and Techniques* 61, 12 (2013), 4678–4688.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[29] Peter Hillyard, Anh Luong, Alemayehu Solomon Abrar, Neal Patwari, Krishna Sundar, Robert Farney, Jason Burch, Christina Porucznik, and Sarah Pollard. 2018. Experience: Cross-Technology Radio Respiratory Monitoring Performance Study. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 487–496.

[30] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[31] Roland Hostettler, Ossi Kaltiokallio, Hüseyin Yiğitler, Simo Särkkä, and Riku Jäntti. 2017. RSS-based respiratory rate monitoring using periodic Gaussian processes and Kalman filtering. In *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 256–260.

[32] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. 2019. VitaMon: measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 1–14.

[33] Fitbit Inc. [n. d.]. Fitbit Charge 3 Advanced Fitness Tracker. https://www.fitbit.com/shop/charge3

[34] Sonal K Jagtap and MD Uplane. 2012. The impact of digital filtering to ECG analysis: Butterworth filter application. In *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*. IEEE, 1–6.

[35] Zhenhua Jia, Amelie Bonde, Sugang Li, Chenren Xu, Jingxian Wang, Yanyong Zhang, Richard E Howard, and Pei Zhang. 2017. Monitoring a person's heart rate and respiratory rate on a shared bed using geophones. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–14.

[36] Faheem Khan and Sung Ho Cho. 2017. A detailed algorithm for vital sign monitoring of a stationary/non-stationary human through IR-UWB radar. *Sensors* 17, 2 (2017), 290.

[37] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[38] T Kondo, T Uhlig, P Pemberton, and PD Sly. 1997. Laser monitoring of chest wall displacement. *European Respiratory Journal* 10, 8 (1997), 1865–1869.

[39] Antonio Lazaro, David Girbau, and Ramon Villarino. 2010. Analysis of vital signs monitoring using an IR-UWB radar. *Progress In Electromagnetics Research* 100 (2010), 265–284.

[40] Changzhi Li and Jenshan Lin. 2010. Recent advances in Doppler radar sensors for pervasive healthcare monitoring. In *2010 Asia-Pacific Microwave Conference*. IEEE, 283–290.

[41] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical human sensing in the light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. 71–84.

[42] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. 2016. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. 164–175.

[43] Xiaolin Liang, Hao Zhang, Shengbo Ye, Guangyou Fang, and T Aaron Gulliver. 2018. Improved denoising method for through-wall vital sign detection using UWB impulse radar. *Digital Signal Processing* 74 (2018), 72–93.

[44] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. 2017. Cardiac scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 315–328.

[45] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 267–276.

[46] Xuefeng Liu, Jiannong Cao, Shaojie Tang, Jiaqi Wen, and Peng Guo. 2015. Contactless respiration monitoring via off-the-shelf WiFi devices. *IEEE Transactions on Mobile Computing* 15, 10 (2015), 2466–2479.

[47] Xiaochen Liu, Pradipta Ghosh, Oytun Ulutan, BS Manjunath, Kevin Chan, and Ramesh Govindan. 2019. Caesar: cross-camera complex activity recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 232–244.

[48] Novelda LLC. [n. d.]. X4M03 Radar Development Kit. https://www.xethru.com/xethru-development-platform.html

[49] Reham Mohamed and Moustafa Youssef. 2017. Heartsense: Ubiquitous accurate multi-modal fusion-based heart rate estimation using smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–18.

[50] Sabikun Nahar, Tuan Phan, Farhan Quaiyum, Lingyun Ren, Aly E Fathy, and Ozlem Kilic. 2018. An electromagnetic model of human vital signs detection and its experimental validation. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 8, 2 (2018), 338–349.

[51] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. ACM, 45–57.

[52] Phuc Nguyen, Xinyu Zhang, Ann Halbower, and Tam Vu. 2016. Continuous and fine-grained breathing volume monitoring from afar using wireless signals. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[53] Van Nguyen, Abdul Q Javaid, and Mary Ann Weitnauer. 2014. Spectrum-averaged Harmonic Path (SHAPA) algorithm for non-contact vital sign monitoring with ultra-wideband (UWB) radar. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2241–2244.

[54] Bingbing Ni, Chi Dat Nguyen, and Pierre Moulin. 2012. RGBD-camera based get-up event detection for hospital fall prevention. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1405–1408.

[55] NONIN. [n. d.]. LifeSense II WIDESCREEN Capnograph and Pulse Oximeter. https://www.nonin.com/products/lifesense2/

[56] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 149–161.

[57] Jaeyeon Park, Woojin Nam, Jaewon Choi, Taeyeong Kim, Dukyong Yoon, Sukhoon Lee, Jeongyeup Paek, and JeongGil Ko. 2017. Glasses for the third eye: Improving the quality of clinical data analysis with motion sensor-based data filtering. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–14.

[58] Neal Patwari, Lara Brewer, Quinn Tate, Ossi Kaltiokallio, and Maurizio Bocca. 2014. Breathfinding: A wireless network that monitors and locates breathing in a home. *IEEE Journal of Selected Topics in Signal Processing* 8, 1 (2014), 30–42.

[59] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. 2010. Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography. *IEEE Transactions on Information Technology in Biomedicine* 14, 3 (2010), 786–794.

[60] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1574–1582.

[61] Ruth Ravichandran, Elliot Saba, Ke-Yu Chen, Mayank Goel, Sidhant Gupta, and Shwetak N Patel. 2015. WiBreathe: Estimating respiration rate using wireless signals in natural settings in the home. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 131–139.

[62] Angel Melchor Rodríguez and J Ramos-Castro. 2018. Video pulse rate variability analysis in stationary and motion conditions. *Biomedical engineering online* 17, 1 (2018), 11.

[63] Budiman PA Rohman, Manjunath Thindlu Rudrappa, Maksim Shargorodskyy, Reinhold Herschel, and Masahiko Nishimoto. 2021. Moving Human Respiration Sign Detection Using mm-Wave Radar via Motion Path Reconstruction. In *2021 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 196–200.

[64] Yu Rong and Daniel W Bliss. 2018. Harmonics-based multiple heartbeat detection at equal distance using uwb impulse radar. In *2018 IEEE Radar Conference (RadarConf18)*. IEEE, 1101–1105.

[65] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. 2018. How does batch normalization help optimization? *Advances in neural information processing systems* 31 (2018).

[66] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[67] Loren Arthur Schwarz, Artashes Mkhitaryan, Diana Mateus, and Nassir Navab. 2012. Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing* 30, 3 (2012), 217–226.

[68] Ghufran Shafiq and Kalyana C Veluvolu. 2014. Surface chest motion decomposition for cardiovascular monitoring. *Scientific reports* 4, 1 (2014), 1–9.

[69] Tetsuya Shimamura and Hajime Kobayashi. 2001. Weighted autocorrelation for pitch extraction of noisy speech. *IEEE transactions on speech and audio processing* 9, 7 (2001), 727–730.

[70] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. IEEE, 1297–1304.

[71] Kuo-Kai Shyu, Luan-Jiau Chiu, Po-Lei Lee, Tzu-Han Tung, and Shun-Han Yang. 2018. Detection of breathing and heart rates in UWB radar sensor data using FVPIEF-based two-layer EEMD. *IEEE Sensors Journal* 19, 2 (2018), 774–784.

[72] Jianxuan Tu, Taesong Hwang, and Jenshan Lin. 2016. Respiration rate measurement under 1-D body motion using single continuous-wave Doppler radar vital sign detection system. *IEEE Transactions on Microwave Theory and Techniques* 64, 6 (2016), 1937–1946.

[73] Raghav H Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-user gesture recognition using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 401–413.

[74] Aditya Virmani and Muhammad Shahzad. 2017. Position and orientation agnostic gesture recognition using wifi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 252–264.

[75] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[76] Dingyang Wang, Sungwon Yoo, and Sung Ho Cho. 2020. Experimental comparison of ir-uwb radar and fmcw radar for vital signs. *Sensors* 20, 22 (2020), 6695.

[77] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. 2016. Human respiration detection with commodity wifi devices: do user location and body orientation matter?. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 25–36.

[78] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. ACM, 65–76.

[79] Xuyu Wang, Chao Yang, and Shiwen Mao. 2015. PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In *IEEE Confer- ence on Computer Vision and Pattern Recognition*. IEEE, 1110–1118.

[80] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 617–628.

[81] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging Multi-Dimensionality for Passive Indoor Wi-Fi Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, 1–16.

[82] Zongxing Xie, Yindong Hua, and Fan Ye. 2022. A Measurement Study of FMCW Radar Configurations for Non-contact Vital Signs Monitoring. In *2022 IEEE Radar Conference (RadarConf22)*. IEEE, 1–6.

[83] Zongxing Xie, Hanrui Wang, Song Han, Elinor Schoenfeld, and Fan Ye. 2022. DeepVS: A Deep Learning Approach For RF-based Vital Signs Sensing. In *In 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22)*. ACM.

[84] Zongxing Xie, Bing Zhou, Xi Cheng, Elinor Schoenfeld, and Fan Ye. 2021. Fusing UWB and Depth Sensors for Passive and Context-Aware Vital Signs Monitoring. In *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 119–120.

[85] Zongxing Xie, Bing Zhou, Xi Cheng, Elinor Schoenfeld, and Fan Ye. 2021. VitalHub: Robust, Non-Touch Multi-User Vital Signs Monitoring using Depth Camera-Aided UWB. In *In the 9th IEEE International Conference on Healthcare Informatics*. IEEE.

[86] Zongxing Xie, Bing Zhou, and Fan Ye. 2021. Signal Quality Detection To-wards Practical Non-Touch Vital Sign Monitoring. In *In the 12th ACM InternationalConference on Bioinformatics, Computational Biology and Health Informatics (BCB '21)*. ACM.

[87] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[88] Jiaming Yan, Hong Hong, Heng Zhao, Yusheng Li, Chen Gu, and Xiaohua Zhu. 2016. Through-wall multiple targets vital signs tracking based on VMD algorithm. *Sensors* 16, 8 (2016), 1293.

[89] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 211–220.

[90] Moustafa Youssef, Matthew Mah, and Ashok Agrawala. 2007. Challenges: device-free passive localization for wireless environments. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. 222–229.

[91] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. 2018. Extracting multi-person respiration from entangled RF signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–22.

[92] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 305–320.

[93] Yi Zhang, Xiuping Li, Rui Qi, Zihang Qi, and Hua Zhu. 2020. Harmonic Multiple Loop Detection (HMLD) Algorithm for Not-Contact Vital Sign Monitoring Based on Ultra-Wideband (UWB) Radar. *IEEE Access* 8 (2020), 38786–38793.

[94] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. ACM, 267–281.

[95] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and Fine-grained Respiration Monitoring via Deep-Learning UWB Radar. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 111–124.

[96] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 313–325.

[97] Bing Zhou, Zongxing Xie, and Fan Ye. 2019. Multi-modal face authentication using deep visual and acoustic features. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[98] Bing Zhou, Zongxing Xie, Yinuo Zhang, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2021. Robust Human Face Authentication Leveraging Acoustic Sensing on Smartphones. *IEEE Transactions on Mobile Computing* (2021).