VitalHub: Robust, Non-Touch Multi-User Vital Signs Monitoring using Depth Camera-Aided UWB

Zongxing Xie*, Bing Zhou†, Xi Cheng*, Elinor Schoenfeld*, and Fan Ye*
* Stony Brook University
{zongxing.xie, cheng.xi, elinor.schoenfeld, fan.ye}@stonybrook.edu
† IBM Research
bing.zhou@ibm.com

***Abstract*—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. This type of data collection, interpretation and evaluation is especially valuable for older adults facing myriads of health challenges. However, respiration harmonics and intermodulation cause strong disturbances to much weaker heartbeat signals, thus robust vital signs monitoring remains elusive. In this paper, we propose *VitalHub*, 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 that achieves near-human performance differentiating signal corruptions from large motion. To monitor multiple cohabiting subjects in-home, we leverage consecutive skeletal poses from the depth data to distinguish between individuals and their activities, providing activity context important to disambiguating critical from normal vital sign variability. 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 reveal how existing techniques for harmonics and intermodulation rely on presumed signal patterns thus may fail under real-world dynamic changes.**

***Index Terms*—Vital signs monitoring, non-touch sensing, longitudinal in-home data collection, aging**

I. Introduction

Basic vital signs including respiration and heart rates are predictors for assessing overall changes in health status, and a myriad of medical conditions including respiratory, cardiac, and sleep [1, 2]. 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, 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 (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.

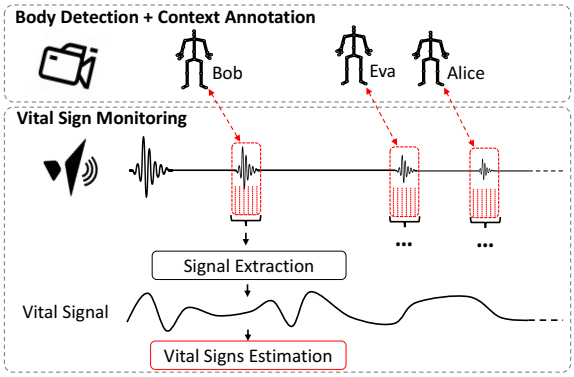


Fig. 1: *VitalHub* leverages depth camera for body detection and context annotation. The location of detected human body is used to segment the UWB signal correspodning to the chest wall of the interested subject, and vital signal is extracted for vital signs estimation.

Recent radio-based passive-sensing solutions [3] exemplified by Wi-Fi [1, 4, 5], FMCW [6], and UWB [7, 8] hold promise for longitudinal in-home monitoring. Temporal and spectral methods extracting vital signs against multipath [5], cluttered environments [9] have been proposed.

However, robustness against harmonics and intermodulation has not received sufficient attention. Because neither heartbeat nor respiratory signals are pure sinusoidal and the human body presents a non-linear channel, high-order harmonics and intermodulations (i.e., linear combinations of RR and HR) exist, and frequently carry energy stronger than the heart rate fundamental frequency. They produce high spectral peaks within the normal heart rate range, e.g., 50–150 beat per minute (bpm). Thus, spectral methods simply pick highest peaks to identify HR easily fail ($1/3$ of the time in our experiments). Making matters worse, we observe that their frequencies and magnitudes are time-varying, and their pattern keeps changing over time, defying simple predictions. In

This work is supported in part by NSF grants 1652276, 1951880, 2028952.
†This work was done when Bing Zhou was with Stony Brook University.

recent radio sensing work, these issues are not described nor appear tackled. The electrical engineering community has completed some studies to address this issue [8, 10], 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 [3, 6] rely on spectral energy or temporal waveform assumptions that are susceptible to dynamic changes, thus not reliable enough.

In this paper, we propose *VitalHub*, a robust vital sign sensing system for longitudinal in-home monitoring using a pair of co-located UWB and depth sensors (as illustrated in Figure 1). 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 RR/HR 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 RR/HR estimates and train a ResNet [11] model to produce a confidence value of how likely the signal is corrupted.

To facilitate longitudinal in-home monitoring in cohabiting scenarios, we leverage the depth sensor to detect human subjects and differentiate their identities from the skeletal walking patterns [12]. Built upon mature pattern recognition methods, VitalHub automatically annotates sensed vital signs with respective context information (e.g., identities and physical activities) for unambiguous vital signs association.

We 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 RR/HR, 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 describe in depth the challenges caused by harmonics and intermodulation, and their serious consequences on heart rate estimation. We do not find sufficient description nor treatment in recent radio sensing work.
- We design a probabilistic weighted framework (PWF) that constantly adjusts the weights in combining the outputs of four estimators based on quantitative evidence of respective patterns, and demonstrate it achieves within 0.3/1.7 bpm error to the upper limit of an idealistic oracle, demonstrating the robustness of PWF.
- We compare three representative methods dealing with harmonics and intermodulation. We show that VitalHub achieves $\leq$5bpm error of HR estimation for 98.5% of the time, while others achieved only 51.2–81.3%. We share insights as to how assumptions they rely on may not hold in reality.
- We compare the heatmap detector against 4 other common methods, and find it achieves near-human performance at 96% for both precision and recall, while others at best 89/83%.

II. Related Work

Three main categories of techniques have been explored for contactless ubiquitous vital sign monitoring: remote PPG [13], acoustic [2, 14], and microwaves including WiFi [1, 5] and other RF based methods [6, 15]. Remote PPG measures the HR from an RGB video of the user's face [13]. It has privacy issues. Reliable data collection may be impacted by skin color, make-up and lighting. Both active [2, 16] and passive [14] acoustic methods leveraging smart speakers or phones have been studied for sensing distance as well as minute chest movement for RR monitoring. However, acoustic methods are limited to near-field monitoring (usually within a distance of 50 cm) due to attenuation. WiFi based methods leverage channel state information (CSI) [1, 5] or received signal strength (RSS) [4]. Other RF methods have exploited techniques including mmWave [15], doppler radar [17], FMCW radar [6], and UWB radar [7]. We observe that respiration harmonics and intermodulation severely disrupt spectral based HR estimation, yet we do not see this described or treated sufficiently in the above work.

The challenges of harmonics and intermodulation is analyzed in [9]. Work from the electrical engineering community [7, 8, 18] has proposed methods based on certain assumptions of the signal's temporal and spectral patterns (e.g., magnitudes between fundamental and harmonic components, gradual changes in HR). We observe such patterns are far from stable, thus these methods often fail. WiBreathe [19] 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 detecting heart rate 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 HR tracking (see §V-C2).

Most radio sensing based work targets quasi-stationary settings. To detect large motions that corrupt signals, many methods use a fixed threshold for phase change or spectrum sharpness [1, 3, 6, 20]. We observe that such fixed thresholds

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

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 (see §V-B).

Measuring and cancelling motion disruptions require extra accelerometers [21, 22], regular RGB cameras or radio sensor pairs [23]. We focus on harmonics and intermodulation in this paper and will explore robust measurement under motion in future work.

III. Design Considerations

A. Hardware Choices for Passive Sensing

Passive sensing does not need any cooperative efforts from users (e.g., charging batteries, wearing devices, or annotating the signal), thus it is critical for longitudinal monitoring in realistic scenarios, especially for older adults with cognitive and physical challenges. To this end, we choose a co-located *UWB* and *depth* sensor pair that complements each other: the UWB signal is sensitive to tiny displacements of the chest wall due to heartbeat and respiration for vital signs extraction; the depth sensor provides context information to help identify the person and segment the UWB signal for further processing.

B. Rationales of Vital Signs Extraction

Heart and respiratory rates are two of the five vital signs collected at each physical examination. It is the combination of heartbeat and respiration that comprises chest displacements. The chest displacement sensed by the UWB sensor can be modeled as:

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

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_r and d_h are the chest displacement amplitudes, and f_r and f_h the rates of respiration and heartbeat, respectively.

The phase modulated by the chest displacement can be modeled as:

$$\phi(t) = \phi_0 + \phi_D(t), \tag{2}$$

where ϕ_0 is the initial phase of the received signal at the nominal distance d_0, and $\phi_D(t) = 2\pi f_c D(t)/c$ is the phase modulated by the physiological movements, f_c is the center frequency of UWB pulse.

C. Robustness Challenges

Robust vital sign extraction based on the derived phase model in (2) 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 (2) needs to be updated as:

$$\phi(t) = \phi_0 + \phi_D(t) + \phi_{STO}(t), \tag{3}$$

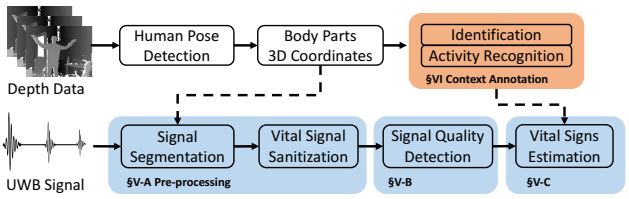


Fig. 2: The overall framework of VitalHub.

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 HRs span a wide range (e.g., 50–150 bpm), the higher order harmonics of respiration can co-exist in the same range as HR. Larger respiration motions also produce strong harmonics, making it difficult to decide the correct fundamental component of HR.

Third, as the realistic channel exhibits non-linearity, the perceived phase due to physiological motion can be more complex than $\phi_D(t) = 2\pi f_c D(t)/c$. The complex non-linear signal can be approximated by its Taylor series as follows:

$$\phi_D(t) = \frac{2\pi f_c}{c}(a_1 D(t) + a_2 D^2(t) + a_3 D^3(t) + \ldots), \tag{4}$$

where a_i is the coefficient of the i-th order term. The higher order terms result in intermodulation products between heartbeat and respiratory signals, manifested as spectral components at frequencies of linear combinations of RR and HR (i.e., $\{mf_h \pm nf_r | m, n \in \mathbb{N}_0\}$). Such components could exist in the normal HR range, making it even more difficult to determine the correct frequency component for heartbeat. The non-linear channel thus coefficients a_i in (4) are time-variant, resulting in unpredictable and dynamic magnitudes of intermodulation components. Therefore, a method that relies on certain assumptions on signal patterns may fail under different patterns.

IV. 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 §III-C for robust vital signs monitoring.

We develop a pre-processing pipeline (in §V-A) to deal with STO issues and extract the vital signal (i.e., the phase change in UWB signal due to physiological motions). We introduce a signal quality detector (in §V-B) 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. We propose a probabilistic weighted framework (PWF) in vital signs estimation (in §V-C) to specifically deal with the challenges in robust HR estimation in the presence of dynamic signal patterns.

While 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). To support unambiguous monitoring in cohabiting scenarios, respective context (e.g., identities and activities) must be correctly associated with the UWB echo pulses from different subjects. We use existing techniques to enable context annotation (in §VI) built upon an existing human pose recognition model [12].

V. 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 RR/HR in presence of unpredictable and dynamic signal patterns.

A. 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. Our UWB sensor has a range of $10\ m$, leading to about 200 range bins. We leverage the human body distance measurement from context annotation module (in §VI) to decide which range bin corresponds to which identified human body, thus further processing signals in those bins. The $5cm$ size is decided based on the amplitude of motion, the penetration effects of signals and errors in distance measurement.

Vital Signal Sanitization. Next we remove the time-variant phase drift $\phi_{STO}(t)$ due to sampling time offset (STO) (analyzed in §III-C). 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:

$$\phi'(t) = \phi(t) - \phi_r(t) = \phi_D(t) + \phi_0 - \phi_0^r, \tag{5}$$

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.

B. 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 [1, 6, 20] 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 RR/HR candidate pair $\{f_r,\ f_h\}$:

$$HM(f_r,\ f_h) = \sum_{z \in \mathbb{Z}(f_r,\ f_h)} A(z), \tag{6}$$

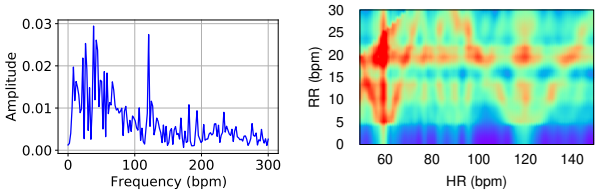
where $A(z)$ denotes the spectral amplitude of the signal at the frequency of z, and $\mathbb{Z}(f_r, f_h)$ is a set of potential harmonic and intermodulation frequencies, which can be expressed as $\{mf_h \pm nf_r | m, n \in \mathbb{N}_0\}$. When the signal is not corrupted much, harmonic and intermodulation frequencies of (f_r, f_h) close to the true RR/HR 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 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". [2]

To learn the spatial-invariant features from the 2-D heatmap, we adopt the ResNet-18 model as the detector. ResNet [11] 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 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 §VIII-A. Signals detected as "available" are passed for vital signs estimation.

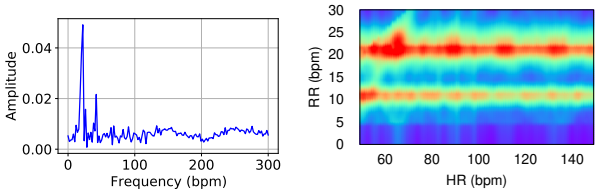
However, signals from the range bin that was directly located by the depth camera may not be suitable for vital signs estimation due to the offset error of the depth measure and the imperfect placement between UWB and depth sensors. We note that adjacent range bins need to be considered to measure vital 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 $\pm 15\ cm$ range) is

[2]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.



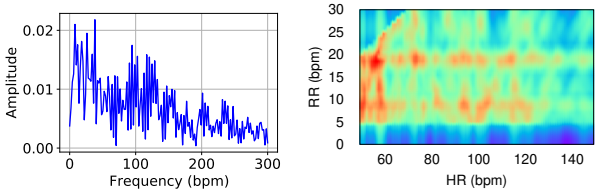
(a) FFT Spectrum. (b) 2-D Heatmap.

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.

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.

C. Vital Signs Estimation

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.

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 §III-C), harmonics and intermodulation from respiration can easily dominate the heartbeat signal and their patterns are dynamic.

To robustly measure the HR, we propose a probabilistic weighted framework (PWF) that 1) incorporates four HR estimators each suitable to one of four identified temporal and spectral patterns; 2) adaptively combines HR candidates generated by the estimators with the quantified cumulative evidence of each pattern; and 3) leverages limits in HR 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, noises behave randomly and can be modeled as Gaussian. We use auto-correlation 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 [5] to extract heartbeat signals because DWT can retain the inherently irregular shape of the vital signals while the conventional filters (e.g, Butterworth filter) 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 (7).

$$\begin{cases} \gamma_k^{(L)} = \sum_{n\in\mathbb{Z}} s[n]\varphi_{2^L n-k}^{(L)}, L \in \mathbb{Z}, \\ \upsilon_k^{(l)} = \sum_{n\in\mathbb{Z}} s[n]\psi_{2^l n-k}^{l}, l \in \{1, \ldots, L\}, \end{cases} \tag{7}$$

where φ denotes the scaling function and ψ the wavelet. The signal can be reconstructed using inverse DWT (8).

$$s[n] = \sum_{k\in\mathbb{Z}} \gamma_k^{(L)} \varphi_{2^L n-k}^{(L)} + \sum_{l=1}^{L} \sum_{k\in\mathbb{Z}} \upsilon_k^{(l)} \psi_{2^l n-k}^{l}. \tag{8}$$

In VitalHub, we select Daubechies(db4) wavelet as the mother wavelet, 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 $\upsilon^{(4)} + \gamma^{(4)}$ (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 for each, including: 1) zero-crossing (ZC); 2) peak interval (PK); 3) local maximum detection in the spectrum of the HR 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 HR 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 HR. PK measures the average interval between adjacent local maxima in a time window, thus HR. 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 [6], LMD detects such high peaks in the HR 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 [8]. Due to partial overlap with the heartbeat fundamental frequency range, sometimes respiration may still produce significant peaks thus erroneous HR estimation.

Using a sliding window, we produce a HR 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 HR estimation as tracking the "trend" of changes, with the state update equation as follows:

$$\hat{x}_t = x_{t-1} + \dot{x}_{t-1}\triangle t + \varepsilon_p, \tag{9}$$

where x_{t-1} is the state (i.e., HR) estimated at time $t-1$, $\hat{x}_t$ is HR 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 HR 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 [8]. The closer a candidate is to any respiration harmonic, the less likely it is true, which can be formulated in the following weight:

$$P_r(c_t^m) = 1 - g_r(\min_{m,n}(abs(c_t^m - n \cdot f_t^r)), \tag{10}$$

  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 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:

$$P_h(c_t^m) = g_h(\min_n(abs(c_t^m - c_t^n)), \tag{11}$$

$$P_h(c_t^n) = g_h(\min_m(abs(c_t^m - c_t^n)), \tag{12}$$

  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. We use an exponential distribution to represent this weight:

$$P_p(c_t^m) = 1 - e^{-\alpha \cdot p(c_t^m)}, \tag{13}$$

  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.* HR is not likely to change abruptly in a short time (e.g., one second), and the next HR is usually close to the current one. Therefore, we quantify how close a candidate is to previous estimation as:

$$P_l(c_t^m) = g(abs(c_t^m - x_{t-1})), \tag{14}$$

  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:

$$\mathcal{L}_t^m = P_r(c_t^m) \cdot P_h(c_t^m) \cdot P_p(c_t^m) \cdot P_l(c_t^m). \tag{15}$$

The normalized weight for a candidate is expressed as:

$$\omega_t^m = \frac{\mathcal{L}_t^m}{\sum_{j=1}^{M_t} \mathcal{L}_t^j}, m = 1, 2, ..., M_t. \tag{16}$$

Then, we take the weighted average of all the candidates as a new measurement:

$$\bar{c}_t = \sum_{c_t^n \in C_t} \omega_t^n \cdot c_t^n. \tag{17}$$

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 [24]). 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:

$$\begin{aligned} K_t &= \frac{\sigma_t^2 \ 1}{\sigma_M^2 + \sigma_t^2 \ 1}, \\ \sigma_t^2 &= (1 - K_t)\, \sigma_{t-1}^2, \\ x_t &= \hat{x}_t + K_t(\bar{c}_t - \hat{x}_t) \end{aligned} \tag{18}$$

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.

VI. Context Annotation

We leverage the skeleton data tracked from the depth sensor as features for user identification and activity context recognition.

User Identification. As the walking pattern is discriminative, we use the consecutive skeleton data in a two-second time window (a few steps when the user enters the monitoring zone) as input for identification. We leverage a deep recurrent model with two stacked standard LSTM layers, each with 128 hidden units, and a fully connected layer with Softmax activation to produce prediction results.

Activity Context Recognition. Since the pre-trained LSTM model has learned sophisticated features from sequential skeleton data, we apply transfer learning by feeding them to a new classifier to recognize activity context.

VII. Testbed

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

A. Implementation

VitalHub uses a COTS IR-UWB sensor XeThru x4m03 [25] as its frontend for wireless sensing. The transmitted pulse is configured to be within the 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 [12] 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. We run the whole pipeline on a 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.

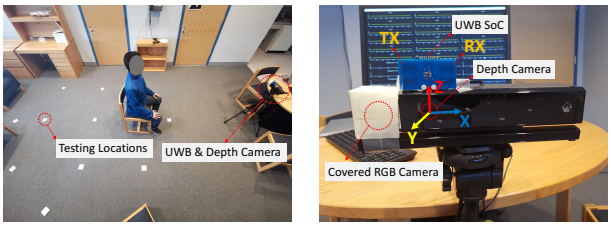
B. Experimental Setup

Figure 6(b) shows the hardware setup of a Kinect XBox One sensor with the *RGB camera covered* and a co-located UWB sensor. We conduct experiments in a room with a size of $4.5 \times 9\ m^2$ (shown in Figure 6(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 [26] and Masimo Pulse Oximeter [27] to obtain heart and respiratory rates as ground truth. Although results for each module are presented separately, VitalHub inherently integrates and produces data in a holistic pipeline concurrently.

VIII. Evaluation

We start with a few microbenchmarks to demonstrate the performance of *signal quality detector* in §VIII-A. Then, we evaluate the end-to-end performance of VitalHub in §VIII-B and §VIII-C. 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.



(a) Testing environment setup. (b) Hardware configuration.

Fig. 6: The experiment environment and hardware setup.

A. 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) [20], moving variance (MVBD) [20], average variance energy (AVE) [1], and flat spectrum (FSD) [6].

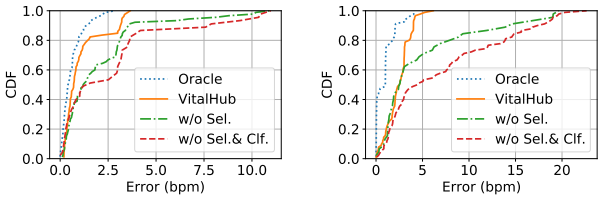
We build a balanced data set consisting of $20,000$ data samples, with equal number of "available" and "unavailable" samples randomly selected from $40,782$ manually labeled ones. Each data sample is the vital signals in a 30-second time window from one of 7 adjacent range bins centered at the depth 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 [28] that minimizes cross entropy as the loss function, which measures the discrepancy between predicted and actual labels.

TABLE I: 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) CDF of RR error. (b) CDF of HR error.

Fig. 7: Boosted performance with signal quality detector.

Table I 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.

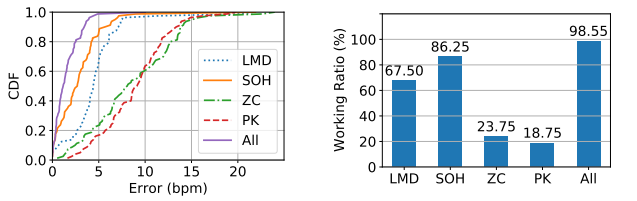
Range Bin Selection. We observe that the distance reported by the depth sensor may not give the range bin with the best signal quality. Thus we search 7 adjacent range bins (± 15 cm) centered at the depth camera reported one, and select the one with the highest signal quality indicator α (provided by 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.

Figure 7 shows obvious performance degradation each time bin selection or classification is removed. With both of them, VitalHub achieves end-to-end RR/HR 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.

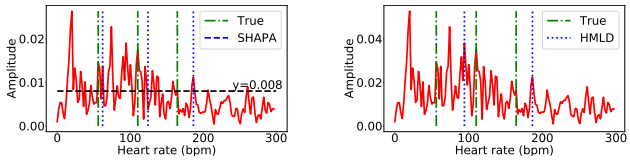
B. Vital Signs Estimators

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 §III-C).



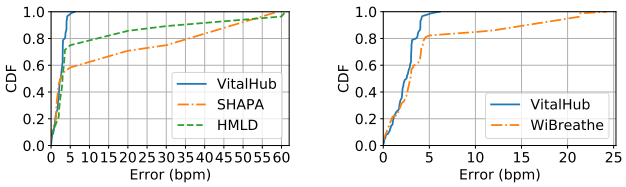
(a) CDF of HR error. (b) Working ratio of different HR estimators.

Fig. 8: Performance of individual estimators.



(a) SHAPA. (b) HMLD.

Fig. 9: Typical spectrum when SHAPA and HMLD both fail.



(a) VitalHub, SHAPA and HMLD on HR estimation. (b) VitalHub and Wibreathe on HR estimation.

Fig. 10: Vital signs estimation under different orientations.

1) Individual Estimators: We evaluate the effectiveness of all the 4 estimators for heart rate estimation (Figure 8). 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 PWF framework combining weak estimators to achieve more robust estimation.

2) Robustness against Harmonics and Intermodulation: We identify two representative methods, SHAPA [7] and HMLD [18], 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 10(a) 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 9(a) shows a typical case where the threshold set at the minimum magnitude of all harmonic peaks. However, many noise peaks exist above the threshold, and some may cause SHAPA fail to locate the true harmonic path. HMLD has similar problems as shown in Figure 9(b). The above shows that algorithms relying on presumed signal patterns are not robust enough.

3) Dealing with Unpredictable and Dynamic Signal Patterns: We implement WiBreathe [19] 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 10 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 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.

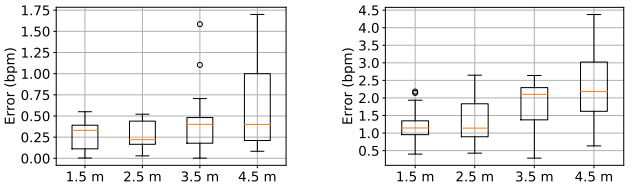
C. User and Environment Factors

We study the impact of user and environment factors to end-to-end vital sign monitoring performance.

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 11. We can see that RR/HR estimations are very stable even at $4.5\ m$, up to the range of the depth sensor.

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 12). Interestingly, we observe that HR accuracy is not affected much by the orientation, but RR error at $90\ degree$ more than triples. The issue of RR being sensitive to the orientation will be discussed in §IX.

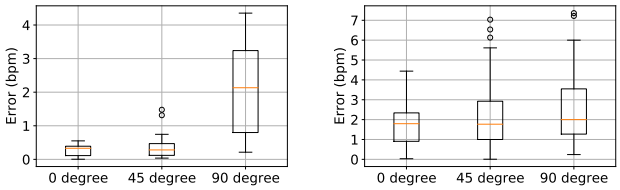
3) Impact of Ambient RF sources: To evaluate the impact of ambient RF sources, we compare the performance in two settings: low Wi-Fi traffic where Wi-Fi signal comes from nearby buildings but no Wi-Fi device running indoors, and intense Wi-Fi traffic where 3 Raspberry Pis, 4 laptops, 4 smartphones and 2 Wi-Fi routers keep streaming data indoors. Figure 13 shows negligible decrease in the accuracy of RR/HR measurement. This 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.



(a) Impact of distances on RR estimation.

(b) Impact of distances on HR estimation.

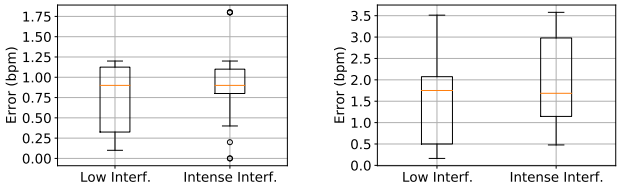
Fig. 11: Vital signs estimation under different distances.



(a) Impact of orientations on RR estimation.

(b) Impact of orientations on HR estimation.

Fig. 12: Vital signs estimation under different orientations.



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

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

Fig. 13: Impact of different ambient RF environment.

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.

IX. 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.

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 HR estimate. If this continues long enough, the PWF may fail to smooth out such erroneous

output and converge to produce some wrong HR estimate. This problem can be alleviated by combining signals in consecutive time windows, but at the cost of reduced recall of the data. We leave it to the future work to find a proper balance.

Sensing Orientation and Range. We observe relatively high errors in RR when the orientation of the subject is near 90 degrees. It is because the chest movement in the mediolateral dimension is much smaller than the frontal dimension, but is comparable to the displacement attributed to the heartbeat. A native but effective solution is to use multi-sensor deployment so the 90-degree orientation of the subject can be avoided by at least one sensor. While the effective range of the current prototype is limited by the UWB and depth sensors we use, it is sufficient for room-size monitoring. We will also explore multi-sensor deployment for scalable coverage.

X. 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. Experiments show that VitalHub achieves performance close to an idealistic but impractical oracle, and we share insights on why existing methods do not handle harmonics and intermodulation well. We believe VitalHub offers a suitable solution for longitudinal in-home vital sign monitoring.

References

[1] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *ACM MobiHoc*, 2015.

[2] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *ACM Mobisys*, 2015.

[3] P. Hillyard, A. Luong, A. S. Abrar, N. Patwari, K. Sundar, R. Farney, J. Burch, C. Porucznik, and S. Pollard, "Experience: Cross-technology radio respiratory monitoring performance study," in *ACM Mobicom*, 2018.

[4] N. Patwari, L. Brewer, Q. Tate, O. Kaltiokallio, and M. Bocca, "Breathfinding: A wireless network that monitors and locates breathing in a home," *IEEE Journal of Selected Topics in Signal Processing*, 2014.

[5] X. Wang, C. Yang, and S. Mao, "Phasebeat: Exploiting csi phase data for vital sign monitoring with commodity wifi devices," in *IEEE ICDCS*, 2015.

[6] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *ACM CHI*, 2015.

[7] V. Nguyen, A. Q. Javaid, and M. A. Weitnauer, "Spectrum-averaged harmonic path (shapa) algorithm for non-contact vital sign monitoring with ultra-wideband (uwb) radar," in *IEEE EMBS*, 2014.

[8] Y. Rong and D. W. Bliss, "Harmonics-based multiple heartbeat detection at equal distance using uwb impulse radar," in *IEEE Radar Conference*, 2018.

[9] A. Lazaro, D. Girbau, and R. Villarino, "Analysis of vital signs monitoring using an ir-uwb radar," *Progress In Electromagnetics Research*, 2010.

[10] R. El-Bardan, D. Malaviya, and A. Di Rienzo, "On the estimation of respiration and heart rates via an ir-uwb radar: An algorithmic perspective," in *COMCAS*, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.

[13] S. Huynh, R. K. Balan, J. Ko, and Y. Lee, "Vitamon: measuring heart rate variability using smartphone front camera," in *ACM Sensys*, 2019.

[14] A. Wang, J. E. Sunshine, and S. Gollakota, "Contactless infant monitoring using white noise," in *Mobicom*, 2019.

[15] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *ACM MobiHoc*, 2016.

[16] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Battracker: High precision infrastructure-free mobile device tracking in indoor environments," in *ACM SenSys*, 2017.

[17] C. Li and J. Lin, "Recent advances in doppler radar sensors for pervasive healthcare monitoring," in *APMC2010*.

[18] Y. Zhang, X. Li, R. Qi, Z. Qi, and H. Zhu, "Harmonic multiple loop detection (hmld) algorithm for not-contact vital sign monitoring based on ultra-wideband (uwb) radar," *IEEE Access*, 2020.

[19] R. Ravichandran, E. Saba, K.-Y. Chen, M. Goel, S. Gupta, and S. N. Patel, "Wibreathe: Estimating respiration rate using wireless signals in natural settings in the home," in *IEEE PerCom*, 2015.

[20] M. Youssef, M. Mah, and A. Agrawala, "Challenges: device-free passive localization for wireless environments," in *ACM Mobicom*, 2007.

[21] H. H. Asada, H.-H. Jiang, and P. Gibbs, "Active noise cancellation using mems accelerometers for motion-tolerant wearable bio-sensors," in *IEEE EMBS*, 2004.

[22] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography," *IEEE T-ITB*, 2010.

[23] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *ACM Mobicom*, 2017.

[24] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *IJEM*, 2012.

[25] X4m03. [Online]. Available: https://www.xethru.com

[26] Nonin. [Online]. Available: https://www.nonin.com

[27] Masimo. [Online]. Available: https://www.masimo.com

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, 2014.