University of California
Santa Barbara

Enabling Novel Sensing Applications with Everyday WiFi Signals

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Belal Salama Amin Korany

Committee in charge:

Professor Yasamin Mostofi, Chair
Professor Joao Hespanha
Professor Upamanyu Madhow
Professor Kenneth Rose

December 2021

The Dissertation of Belal Salama Amin Korany is approved.

Professor Joao Hespanha

Professor Upamanyu Madhow

Professor Kenneth Rose

Professor Yasamin Mostofi, Committee Chair

December 2021

Enabling Novel Sensing Applications with Everyday WiFi Signals

by

Belal Salama Amin Korany

To Rana, Mariam, and my Parents.

Acknowledgements

First, I would like to thank my advisor, Prof. Yasamin Mostofi, for all her guidance throughout my PhD. Her constant support has always succeeded in making me go for the extra mile, and achieve results I had not imagined I could ever achieve. I would also like to thank Prof. Joao Hespanha, Prof. Upamanyu Madhow, and Prof. Kenneth Rose for serving on my committee, and for all their constructive input on my work.

I spent most of my time here at UCSB working in our lab alongside my labmates: Chitra, Herbert, Saandeep, Arjun, and Anurag. I enjoyed each and every moment of working with them, and for that I am extremely grateful. Thanks shall also go to the rest of my friends in Santa Barbara: Mohamed Abdelghany, Ahmed Wahba, Ahmed Elshafiy, Ahmed Ahmed, and Ali Farid, for making my stay in Santa Barbara much more enjoyable.

Finally, I am forever grateful for my parents and my wife Rana. For without their love and constant support, I would not have been able to achieve my current standing.

Curriculum Vitæ

Belal Salama Amin Korany

Education

2021	Ph.D. in Electrical Engineering (Expected), University of California, Santa Barbara.
2015	M.Sc. in Electrical Engineering, Cairo University, Egypt.
2012	B.Sc. in Electrical Communications Engineering, Cairo University, Egypt

Publications (* equal contribution)

- **B. Korany**, and Y. Mostofi, "Nocturnal Seizure Detection Using Off-the-Shelf WiFi," in the IEEE Internet of Things (IoT) journal, 2021.
- **B. Korany**, and Y. Mostofi, "Counting a Stationary Crowd Using Off-the-Shelf WiFi," in ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), June 2021.
- **B. Korany**, H. Cai, and Y. Mostofi, "Multiple People Identification Through Walls Using Off-the-Shelf WiFi," in IEEE Internet of Things (IoT) journal, vol. 8, no. 8, pp. 6963-6974, April 2021.
- H. Cai*, **B. Korany***, C. R. Karanam*, and Y. Mostofi, "Teaching RF to Sense without RF Training Measurements," in ACM Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT), vol. 4, no. 4, Dec. 2020.
- **B. Korany***, C. R. Karanam*, H. Cai*, and Y. Mostofi,"XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage," in ACM Int. Conf. on Mobile Computing and Networking (MobiCom), 2019.
- C. R. Karanam, **B. Korany**, and Y. Mostofi,"Tracking from One Side – Multi-Person Passive Tracking with WiFi Magnitude Measurements," in ACM International Conf. on Information Processing in Sensor Networks, (IPSN), 2019.
- **B. Korany**, S. Depatla, and Y. Mostofi, "Subspace-Based Imaging Using Only Power Measurements," in IEEE Sensor Array and Multichannel Signal Processing Workshop, 2018.
- **B. Korany***, C. Karanam*, and Y. Mostofi, "Adaptive Near-Field Imaging with Robotic Arrays," in IEEE Sensor Array and Multichannel Signal Processing Workshop, 2018.
- C. R. Karanam*, **B. Korany***, and Y. Mostofi, "Magnitude-based Angle-of-Arrival Estimation, Localization, and Target Tracking," in the ACM International Conference on Information Processing in Sensor Networks (IPSN), 2018.

Abstract

Enabling Novel Sensing Applications with Everyday WiFi Signals

by

Belal Salama Amin Korany

Due to the recent rapid growth of the number of wirelessly-connected devices, wireless signals are everywhere these days. These signals interact with our surroundings, e.g., by reflecting off of the objects/people in the environment, thereby carrying crucial information about them. Extraction of such information then facilitates many applications in surveillance, security, and smart homes, to name but a few.

In this dissertation, we investigate the possibilities of utilizing wireless signals of off-the-shelf devices (e.g., WiFi) to enable various novel sensing applications. In the first part of the dissertation, we develop a tool that converts video footage to WiFi signals. More specifically, given a video footage of a person engaged in some activity, our proposed tool generates/simulates the WiFi signals that would have been measured if a WiFi device was placed near the person in the video. This opens up the door to a wide range of new possibilities in RF sensing. We showcase the importance of our proposed tool in two sample applications. First, we develop a through-wall person identification system that can use the WiFi power measurements when a person is walking behind wall in a WiFi-covered area to determine if this person is the same as the one walking in a given video clip. Second, we show that our proposed tool can greatly reduce the training burden of learning-based sensing systems, by generating a training dataset from the vast available video data, and without the need for the laborious and time consuming process of real training WiFi data collection. As a case study, we train a gym activity classifier, for the first time, solely based on videos of people performing different gym exercises, and

without collecting any real training data.

In the second part of the dissertation, we discuss how WiFi signals can be used for nocturnal seizure detection in a robust, fast, and contactless manner, which will provide much better support for epilepsy patients and their caregivers. We develop a new mathematical characterization for the received WiFi signals during different types of nocturnal sleep motions: breathing, normal movements, and seizures. We then propose a WiFi processing pipeline that detects all non-breathing motions and classifies whether they are seizures or not.

Finally, we discuss how to count a stationary (seated) crowd using off-the-shelf WiFi, based on the natural in-place motions that people naturally engage in while seated (also called fidgets), such as changing their pose or crossing their legs. We develop a mathematical model, inspired from Queuing Theory, that relates the fidgeting statistics of a crowd to the number of people in this crowd. Based on this modeling, we propose a WiFi processing pipeline that extracts the fidgets of the crowd and estimates the number of people accordingly. We experimentally validate all of our developed algorithms with several test subjects in several different environments with different levels of clutter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent years have witnessed a rapid growth in the number of wirelessly-connected devices, due to the emergence of new technologies such as the Internet of Things (IoT) and smart homes. This has resulted in the ubiquity of wireless signals, such as WiFi. As such, researches have expressed great interest in utilizing these signals beyond communications, notably for sensing and learning about the environment. More specifically, these Radio Frequency (RF) signals interact with different objects in the environment by reflecting off of their surfaces, thereby carrying crucial information about the surroundings. Extraction of such information from the wireless signals (i.e. wireless sensing) results in an environmental awareness that opens the door for a wide space of new possibilities and applications. For instance, WiFi (which is one of the most widespread forms of wireless connectivity) has been utilized for person identification [1, 2, 3, 4, 5, 6], occupancy estimation [7, 8, 9, 10, 11, 12, 13, 14], health monitoring [15, 16], among other applications.

In this dissertation, we push the frontiers of sensing with wireless signals by proposing and implementing novel sensing applications that have not been enabled before. In particular, this dissertation focuses on the following novel wireless sensing applications:

- **Gait-based Person Identification:** In the first part of the dissertation, we explore the possibility of WiFi-based person identification (based on their walking patterns) from candidate video footage. More specifically, we show that it is possible for WiFi to determine if a person walking in an area is the same as the one in a given video footage. This identification across modalities (i.e. video to WiFi) has not been achieved before, to the best of our knowledge. Moreover, we propose a framework for simultaneous *multiple people* identification.

- **Teaching RF to Sense without RF Training Measurements:** In the second part of the dissertation, we build upon our initial exploration of the interplay between vision and RF sensing (mentioned above) and propose a framework to train any RF-based sensing system without collecting RF training measurements. More specifically, we show that it is possible to use the vast already-available online video data to generate instant RF signals for the purpose of training learning-based sensing systems. As a sample case study, we train a classifier that recognizes 10 different gym exercises using only available videos of people performing these exercises. To the best of our knowledge, this is the first time that an RF sensing system is trained using only video footage, without any RF training data collection.

- **Nocturnal Seizure Detection:** In the third part of the dissertation, we develop a new mathematical model and a resulting system design for seizure detection using WiFi signals, which is the first of its kind, to the best of our knowledge. More specifically, we propose a framework to process the WiFi signals measured on a device placed near a sleeping epilepsy patient, in order to detect whether he/she is having a seizure or not. We test our proposed system on seizures simulated by 20 different actors in 7 different environments, showing a seizure detection rate of 93.85% and a probability of false alarm of 0.0097.

- **Counting a Stationary Crowd:** In the last part of this dissertation, we propose a framework that enables WiFi to count the number of people in a stationary (seated) crowd. In contrary to the traditional WiFi-based crowd counting which relies on people to constantly move, stationary crowd counting is a more challenging problem that has not been solved before, due to the lack of major body motions. To solve this problem, we show that the aggregate natural fidgets of a crowd carries crucial information on the crowd count. We then develop a new mathematical model that describes the collective fidgeting behavior of a stationary crowd and explicitly relates it to the total number of people, by borrowing tools from Queueing Theory. We experimentally validate our approach by counting up to and including 10 people in 4 different environments, showing a high counting accuracy of 93.6%.

It is noteworthy that we develop the aforementioned applications with only off-the-shelf WiFi devices, which makes the problem considerably challenging. Furthermore, all of our proposed solutions do not utilize the absolute phase measurements of the WiFi receiver, since commodity WiFi devices suffer from several sources of phase synchronization error [17].

We next introduce the aforementioned topics in more detail, place them in the context of the state-of-the-art by reviewing the literature in the respective domains, and summarize our contributions in each of these areas.

1.1 Gait-based Person Identification

In the first part of the dissertation (Chapter 2), we propose a novel WiFi-video cross-modal person identification system, which we call XModal-ID (pronounced: Cross-Modal-ID). More specifically, given WiFi measurements of an unknown person walking in an unknown area, and the video footage of some walking person, XModal-ID determines

whether it is the same person in both the WiFi area and the video footage. One key characteristic of XModal-ID is that it does not require any prior wireless or video data of either the person to be identified or the area where the identification is to be conducted. It can further work through walls. To the best of our knowledge, such a cross modal gait-based identification system has not been studied before. This new technique, which draws from both areas of vision and RF, can enable a wide range of new real-world applications that would not be possible with existing technologies, especially in security and surveillance, and personalized service provisioning.

Furthermore, as opposed to all existing person identification systems which focus on single person identification, in Chapter 3, we extend our person identification system to identify *multiple* people simultaneously behind walls.

1.1.1 Related Work

WiFi-based Person Identification: Recently, there has been considerable interest in using off-the-shelf WiFi devices for gait-based person identification. WiFiU [1] uses WiFi CSI to generate spectrogram-based gait features, which are then used to classify the identities of a pre-defined set of people. WiWho [2] uses the time-domain signals measured during people's motion to identify people. Similarly, a few other papers [3, 4, 5, 6] identify a person from a priorly-known set of people. In addition to walking, Wang et al. [18] show that respiration patterns can also be used for identification. WiID [19] uses the CSI measured while a person performs several actions for identification, using two links in the area. Shi et al. [20] identify a person based on his/her daily habits. All these existing approaches require the transceivers to be in the same area as the person, with a line-of-sight view at all times, with the only exception of Hoble [21], which uses a Software Defined Radio (SDR) to identify people in both line-of-sight and through-wall

settings in a known area.

All these existing RF-based papers identify people from a pre-defined group and require prior wireless measurements of these people for training. In other words, they cannot handle new people without retraining. They also require the training and test walking paths/actions and locations to be the same. Thus, a model that is trained in one location and on one type of path cannot be used in other scenarios. The radar-based approaches further require extensive hardware setup. Moreover, aside from [21], none of the existing methods have through-wall identification capabilities. In this dissertation, on the other hand, we propose a novel person identification system that does not require training with prior measurements of the subjects/areas, does not require the test areas/tracks to be known, and can identify people through walls. Finally, our proposed system enables a new set of applications not possible before, i.e., given a video footage of a person, it can detect if this person is present in a WiFi area.

It is noteworthy to mention that all the existing gait-based identification systems are only applicable to situations where there is only one person walking in the WiFi area, and will fail when multiple people are simultaneously walking.

Video-based Person Identification: Video-based person identification using gait is a well-studied problem in the computer vision literature. There are broadly two types of approaches: model-based, where gait features are extracted by fitting a walking human model to the video frames, and model-free, where features are extracted directly from the video. Model-based methods include fitting a walking human with ellipses [22], estimating the lengths of body parts and joint angles [23], and estimating the joint trajectories [24]. Model-free approaches rely on the person's silhouette in the video. Commonly-used features include the silhouette key frames [25] and gait energy image [26]. These features are then fed into machine learning pipelines for training. We refer the readers to [27] for a detailed survey. Overall, video-based methods require installing cameras everywhere

and lack through-wall identification capabilities.

1.1.2 Contributions

Our main contributions in this part of the dissertation (Chapters 2 – 3) are as follows:

1. Given a video footage of a person engaged in some activity, we develop a new tool to simulate the WiFi signals that would have been measured if the person in the video was near a pair of WiFi transceivers. More specifically, we extract a 3D mesh model of the person in the video, use eigen-analysis to align the person to an arbitrary coordinate system, and apply Born approximation to simulate the corresponding WiFi CSI magnitude measurements if the person in the video was in a WiFi-covered area.

2. We use the video-to-WiFi conversion tool to generate the WiFi signals of people walking in videos, and develop XModal-ID: a WiFi-based person identification system that captures the gait characteristics of a person based on WiFi CSI magnitude signals. More specifically, we utilize a combination of Short-Time Fourier Transform and Hermite functions to generate a *spectrogram* of the WiFi signal, and extract a key set of features that are subsequently used for identification. We further propose a way to extract key parts of the spectrogram as well as the direction of motion, which allows us to do identification, without the need to know the track of the person.

3. We extensively evaluate our person identification framework using a large test set, where all the test subjects and test areas are completely unknown in the training phase, thus allowing us to demonstrate the generalizability of XModal-ID to new, unknown people and environments. In the test set, there are 8 subjects, 2 video

areas, and 5 WiFi areas, including 3 areas where the transceivers are placed behind a wall and scenarios where the walking paths are complex. The walking paths are further assumed unknown in all the experiments. Overall, the test set contains a total of 2,256 pairs of WiFi and video samples to be identified. Given a pair of video and WiFi samples, XModal-ID achieves a binary classification accuracy of 85%, in judging whether the two samples belong to the same person.

4. We extend our person identification system to identify multiple people walking simultaneously in a WiFi-covered area. More specifically, given the received WiFi magnitude measurements on a receiver with a small antenna array, we first separate the received signals from each individual subject based on their angles-of-arrival to the receiver array. Then we extract the gait features from each of the separated signals to identify each subject separately. We extensively test this approach with 92 experiments where 2 or 3 test subjects (randomly selected from a pool of 6 subjects) walk simultaneously in one of 4 different test areas. Our proposed approach results in an identification accuracy of 82%.

1.2 Teaching RF to Sense Without RF Training Measurements

In this part of the dissertation (Chapter 4), we propose to extend our video-to-WiFi conversion tool (first discussed in Chapter 2) to generate instant RF signals that can be used to train any learning-based RF sensing system. The motivation for this is that learning-based sensing systems need a large number of RF measurements for training, which requires a laborious prior RF data collection process. Moreover, in spite of the large-scale training measurements, the performance of these systems degrades

considerably when operating with a new setup (e.g., new transceiver placement) or in an area that differs from the setup/area of the training phase [28, 29]. Furthermore, for classification-related tasks, existing systems are not scalable, as they cannot be used for classes not seen during training, requiring additional RF training measurements for the new classes. On the other hand, there are many freely-available vision datasets pertaining to many different activities these days. In this chapter, we show how how we can translate such vision datasets to instant RF datasets for training motion-based RF sensing systems.

As a case study for this idea, we use our video-to-WiFi conversion tool in a realistic WiFi sensing application of gym activity classification. In this case study, we train a classifier using only the WiFi data generated from YouTube videos of people performing the gym activities. We then extensively test the trained classifier with real WiFi data of people performing the gym activities in different areas, and show that it can classify the activities with a high accuracy. To the best of our knowledge, this is the first time that a real-world RF sensing system is enabled with only video-based training data, and without any real RF measurements for training purposes.

1.2.1 Related Work

In recent years, there has been a large body of work that uses RF signals for sensing human motion, activity, and behavior, in order to enable various useful applications, such as person identification, vital signs detection, fall detection, and activity recognition [30, 31, 32, 33, 34]. In terms of activity recognition, great progress has recently been made towards enabling device-free RF sensing. Several papers have utilized specialized hardware or radar for activity recognition. For instance, [35, 36, 37] use USRP and/or radar to classify daily human actions. [38] uses an FMCW radar to capture the human

pose. There has further been considerable interest in utilizing off-the-shelf WiFi devices to perform human activity recognition. For instance, CARM [29] utilizes the relationship between the WiFi signal variations and the speeds of human body parts for classifying a set of 8 activities. [39] achieves activity recognition through-walls on a set of 7 daily activities. [40, 41] utilize deep learning to classify various daily activities. In terms of classifying gym activities, such as push-up and jumping jack, there is a limited number of work [42, 43, 28] that have used RF sensing systems to enable this application, for sets of 4, 9, and 10 physical exercises, respectively, by relying on line-of-sight crossing (i.e., blocking the direct path between the Tx and Rx).

All of these state-of-the-art methods on different aspects of activity recognition, however, need to collect a large amount of prior RF measurements for training purposes. Such data collection typically involves asking several people to perform all the designated actions in different areas. Furthermore, the collected training data is based on a specific experimental setup and a specific set of activities, which cannot be reused for other setups or applications with different activities. [38] has further collected real video measurements of the scene, simultaneous with RF training measurement collection, in order to annotate the collected RF data for the purpose of pose estimation. [44] has recently released their collected RF training dataset for WiFi-based action recognition. While releasing collected data could be useful for the research community, such a dataset is heavily dependent on the activities involved and is not generalizable to other activities/behaviors. Furthermore, the released data is heavily dependent on the sensing setup, such as the number of transceivers, their placement, and the frequency of operation. Thus, this dataset cannot be used for a sensing system with a different configuration.

1.2.2 Contributions

Our main contributions in this part of the dissertation (Chapter 4) are as follows:

1. We propose a novel idea: to train a human-motion-related RF sensing system without any training data, but through leveraging the vast amount of online available videos. More specifically, we propose a way to translate the massive readily-available online videos of people's motions and activities to instant RF data and use them to train RF sensing systems. This idea eliminates the labor-intensive process of collecting RF measurements for training, is **scalable** and is **applicable to any motion-based RF sensing system**. It further allows for easy re-training when new classes and/or a different RF sensing setup are given. Finally, it can enable new possibilities beyond system design and towards system analysis, in order to understand the fundamental limitations of a particular RF sensing system, as a function of the underlying activities, the amount of given resources, and the setup. Overall, researchers can utilize this idea to train RF sensing systems with no RF data collection.

2. Our pipeline consists of the following steps. In order to build a classifier pertaining to a number of motion-related activities, we first gather several online videos of different instances of the activities of interest. We then utilize a state-of-the-art vision-based human shape reconstruction algorithm to build a 3D mesh of the person in each video, as a function of time. Here we have to address a number of challenges. The person in an online video, for instance, could have been captured from any viewpoint, resulting in an unknown coordinate system, or could be doing a variety of different motion-related activities. Our proposed 3D mesh alignment technique via eigen-analysis then enables proper mesh extraction and positioning. We then simulate the corresponding RF signal that would have been measured if

this person was in an RF-covered area, by modeling the interaction between the extracted human mesh and the electromagnetic waves propagating in the environment. In doing so, we take into account the needed sensing setup, e.g., the locations of the transceivers relative to the person and the frequency of operation. Once we translate all the videos to the RF domain, we perform time-frequency analysis on the generated RF data and extract key features from the corresponding RF spectrograms. We then use the extracted features to train a neural network to classify the underlying activities. This RF sensing system that is solely trained on available online videos is then ready for testing in any RF area/setting.

3. We demonstrate the efficacy of our proposed video-based training framework by implementing it for a realistic WiFi sensing application of gym activity classification using only WiFi CSI magnitude measurements of a small number of links. In this case study, we consider 10 different gym activities and generate a corresponding WiFi training dataset only from YouTube videos of people performing these activities. We then use our pipeline to train a WiFi-based gym activity classifier using only the video dataset. We extensively test the trained system in 3 real-world WiFi test areas, where 10 subjects are recruited to perform the activities. We achieve an accuracy of 86% in correctly classifying the gym activity when using a small period of the activity, which may contain a few repetitions of the same exercise (5.1 on average), and an accuracy of 81% when only considering one repetition. Overall, this demonstrates that our proposed idea can eliminate the labor-intensive collection of RF training measurements and enable training RF sensing systems with only already-available video data. It further shows the first demonstration of WiFi-based gym activity classification without any RF training data.

1.3 Nocturnal Seizure Detection

Epilepsy is a neurological disorder that causes a patient to have different kinds of seizures. Epilepsy is treated using different Anti-Epileptic Drugs (AEDs), depending on the specific type of seizure it is causing. Assessment of the ongoing seizure treatment requires the caregivers of the patient to continuously monitor and document the seizures (i.e., their frequency and duration). Seizures which take place during night sleep (**medically known as *Nocturnal Seizures***) then pose a higher risk for epilepsy patients and can be dangerous, since they can go unobserved by the caregivers [45]. This necessitates the need for in-home seizure monitoring devices that can detect nocturnal seizures in epilepsy patients and alert their caregivers.

In this part of the dissertation (Chapter 5), we propose to utilize RF signals to detect nocturnal seizures in epilepsy patients. Using everyday RF signals, i.e. WiFi signals, for such a task has several advantages. First, it is an affordable solution when compared to the high cost of existing approaches. It is also contactless since it does not require the patient to wear any device or have units installed under their mattress. Moreover, an RF-based system, unlike cameras, does not require any lighting conditions to accurately achieve its task. In this dissertation, we then propose to use a pair of WiFi transceivers to detect nocturnal seizures. More specifically, we propose a *robust*, *fast*, and *theoretically-driven* approach to process the WiFi Channel State Information (CSI) measured on a WiFi receiver device placed near a sleeping patient, in order to extract their motion information and decide whether the motion indicates a seizure or not. By "robust", we mean that our proposed framework has a very low probability of false alarm, i.e., it has a very low probability of declaring a seizure when there is none, while detecting all the seizures with a high probability. This is important as the sleeping person may have several normal body movements, such as pose adjustments, and they

should not be classified as a seizure. By "fast", we mean that our system detects a seizure in a very short time since its onset, in order to alert the caregiver in a timely manner. Finally, by "theoretically-driven", we mean that our proposed approach is backed by a new and rigorous mathematical characterization of the spectral content of the received signal during sleep-related movements: seizure, normal body movements, and breathing.

1.3.1 Related Work

To the best of our knowledge, this work is the first to use RF signals for seizure detection. In this section, we summarize the state-of-the-art related to different aspects of our problem of interest.

WiFi-based Vital Signs Monitoring: There has been a great body of work on utilizing wireless signals for vital signs monitoring, e.g. using high-bandwidth radar [46], mmWave [47], or WiFi. In this dissertation, we are interested in utilizing off-the-shelf WiFi devices for seizure detection.

Several papers have utilized the fine-grained WiFi CSI magnitude data for breathing rate and/or heart rate estimation [16, 48]. Other researchers utilized the CSI phase difference between receiver antennas to achieve the same task [49, 50]. None of such RF-based existing work, however, is on seizure detection. Nevertheless, our findings can have a significant impact on such work for the following reason. All the existing WiFi CSI-based breathing rate estimation work assume that the bandwidth of the received CSI signal, in the vicinity of a person who is breathing normally, is the same as the breathing rate. In order for us to develop a robust nocturnal seizure detection system, we also need to fundamentally understand and characterize the spectral content of normal breathing. As we shall see, using our proposed rigorous mathematical analysis, the bandwidth of the WiFi signal caused by normal breathing is not necessarily the same as

the breathing rate and can be higher. As such, this dissertation can contribute to the ongoing research that is using breathing signals for other health monitoring applications. In fact, our mathematical analysis can immediately explain the observation made in [48] that the quality of the breathing rate estimation, which was designed assuming the signal bandwidth is the same as the breathing rate, degrades at some locations. Similarly, it can explain the unexplained frequency peaks that were observed in [51] and were attributed to noise.

Seizure Detection and Analysis: In-home seizure detection is an important topic that has gained a lot of attention in the research community. Most seizure detection algorithms in the literature rely on the detection of the motion of the clonic phase of the tonic-clonic seizure via accelerometry [52, 53, 54], and/or video analysis [55, 56]. In accelerometry, a wearable accelerometer is attached to one or more of the patient's body parts, such as wrist, ankle, and/or chest. In addition to their high cost, wearable devices are usually not well tolerated by certain groups of patients, such as children and people with intellectual disabilities, who usually try to dislodge the devices [56], and wearing them during sleep can be even less desirable. Furthermore, the authors of [57] concluded that commercial wrist-worn watches have a seizure detection accuracy of 89.7%, which is not very high. Video-based seizure detection has shown a good detection accuracy of more than 95%, but with a high false alarm rate of 0.78 events per night [56, 55]. However, video-based detection requires a clear unobstructed view of the patient with good lighting conditions, which may not always be possible, and further invade the patient's privacy. Other studies have focused on the evaluation of commercial seizure detection products. For instance, [58] reported that the MP5 mattress units [59] have a detection accuracy of only 75%, and that several patients have reported discomfort when such units are installed under their mattress. Overall, an accurate, non-invasive, comfortable, and affordable way of detecting nocturnal seizures is lacking, which is the

main motivation for our work.

1.3.2 Contributions

1. We develop a novel and rigorous mathematical model for the received CSI squared magnitude signal as well as the CSI antenna phase difference during different kinds of motions relevant to sleep: seizure, normal body movements, and breathing. More specifically, we first show that both the WiFi CSI squared magnitude and phase difference signals are frequency-modulated by the body motion. We then show our main theoretical contribution: to mathematically characterize the spectral content/bandwidth of the WiFi CSI signal during the aforementioned motions. Based on this new spectral analysis, we then show that the bandwidth of the received WiFi signal can be used to robustly and efficiently differentiate seizure events from normal sleep movements.

2. Based on our theoretical analysis, we propose a new pipeline for the detection of nocturnal seizures using WiFi CSI, which consists of the following 3 steps. First, our data pre-processing pipeline denoises the raw measured CSI and selects the least noisy data streams of different receiver antennas/subcarriers using our spectral analysis findings. Then, our event detection algorithm detects any kind of non-breathing motion, based on the spectral content of the denoised CSI. Finally, an event classification algorithm decides whether a detected event is a seizure or normal body movement, based on the bandwidth of the WiFi signal during the event.

3. In order to validate our proposed framework, we carry out extensive experiments on 20 test subjects (5 females and 15 males) in 7 different locations of typical bedrooms, where the subjects act out seizures and normal sleep movements while we collect WiFi CSI data. In total, we collect 260 different seizure instances and

410 different normal non-breathing sleep movement instances. Our system was able to detect 93.85% of the seizures with an average response time of 5.69 seconds since the onset of the seizure, which is much less than the state of the art. Moreover, in terms of false alarm rate (the probability that a normal sleep event is classified as seizure), our proposed framework had a false alarm probability of 0.0097, which indicates its robust performance. We further study the impact of varying several different parameters (e.g., TX/RX positions) on the performance of our proposed system. Overall, our results establish that our proposed mathematically-motivated system is fast and robust and is also independent of person's pose/orientation.

As we shall see, our derivations can also contribute beyond seizure detection, in the general area of breathing-based RF sensing, since they show that a common assumption regarding the frequency content of the received signal during normal breathing is not always correct, explaining some of the unexplained observations in the corresponding literature.

1.4 Counting a Stationary Crowd

Occupancy estimation and crowd counting have gained a considerable attention in the RF sensing literature due to their relevance to different applications [7, 9, 13]. While there has been a great body of recent work on crowd counting using WiFi signals, most have considered counting mobile people, i.e., people have to walk around to be counted [7, 11, 12]. There are, however, many real-world scenarios where it is of interest to count a seated crowd where individuals are not moving around, such as the attendees of an event/seminar or readers in a library. In the final part of the dissertation (Chapter 6), we are interested in crowd counting when the crowd is *stationary* (i.e. seated), using a pair of off-the-shelf WiFi transceivers.

1.4.1 Related Work

Counting a Mobile Crowd: A number of recent papers have tackled the problem of counting a mobile crowd where each person is moving. In these papers, it is assumed that every person in the group is walking in the WiFi coverage area. These counting methods can be broadly divided into two categories: model-based methods [7, 8, 9], and learning-based methods [11, 12, 13, 14, 10]. In model-based methods, the counting is based on a mathematical modeling of the received WiFi signal based on people's motion. On the other hand, learning-based methods rely on training a neural network with several raw WiFi signals (or manually-extracted features thereof) collected when different number of people are present in the WiFi area. Thus, they can only be used in the area/configurations that they were trained on. Overall, the methods developed for counting a mobile crowd are not applicable to our problem of counting a stationary crowd.

Counting a Stationary Crowd: Due to its difficulty, stationary crowd counting has not gained as much attention in the literature. A couple of approaches have been proposed to utilize breathing signals in order to count the people [60, 51]. Breathing-based approaches, however, suffer from two main drawbacks. First, the crowd is required to stay still, without any motion, for an extended period of time in order to measure their breathing signals, which is an overly restrictive and impractical assumption. Second, breathing-based methods rely on different people to have different breathing rates in order to differentiate and count them, greatly limiting the total number of people they can count.

The authors of [61] trained a deep neural network for counting stationary people. However, they had to collect extensive training data pertaining to several different number of people, seated in various configurations in the test area. Furthermore, such

learning-based approaches can only count the number of people that they are trained on, and with the people in the same seating configuration as in training, and in the same environment, and their performance degrades considerably when deployed with unseen configurations, even in the same environment [61].

1.4.2 Contributions

1. We propose that the aggregate fidgets of the crowd carry crucial information on the crowd count. We then develop a new mathematical model that describes the collective fidgeting behavior of a stationary crowd, and explicitly relates it to the total number of people. More specifically, we mathematically characterize the distribution of the *Crowd Fidgeting Periods (CFPs)*, which we define as the time durations in which at least one person in the WiFi area is fidgeting, as well as that of the *Crowd Silent Periods (CSPs)*, which we define as the time durations in which none of the people are fidgeting, and show their dependency on the total number of people. Based on this analysis, we then propose a Maximum A Posteriori (MAP) estimator of the number of people, using our derived mathematical models of CFPs and CSPs.

2. In developing our mathematical models, we reveal how our problem of interest resembles a several-decade-old $M/G/\infty$ queuing theory problem. More specifically, we show in details how the CSPs are similar to the times when no customer is at a queue that has infinite servers, while the CFPs resemble the times when at least one customer is being served at such a queue. We then explicitly characterize the similarity between the two problems, which allows us to borrow mathematical tools from a 1985 paper on $M/G/\infty$ queues.

3. We develop a framework for processing the WiFi Channel State Information (CSI) measured at a WiFi receiver, in order to extract the times in which at least one person

in the coverage area is fidgeting (i.e. CFPs). More specifically, we show how the received WiFi CSI bandwidth can be used to indicate whether at least one person in the WiFi area is fidgeting, which can then be used to detect the CFPs and CSPs.

4. We extensively validate our proposed approach with a total of 47 experiments in four different environments (including through-wall settings), in which up to and including $N = 10$ people are seated and behaving normally, while a pair of WiFi transceivers, located on one side of the area, make CSI measurements. In total, 27 different subjects participated in our experiments. We test our system in different scenarios representing various engagement levels of the crowd, such as attending a lecture/presentation, watching a movie, and reading. We further test our proposed system with different number of people seated in several different configurations. Our evaluation results show that our proposed approach achieves a very high counting accuracy, with the estimated number of people being only 0 or 1 person off from the true number of people 96.3% of the time in non-through-wall settings, and 90% of the time in through-wall settings. Overall, our results show the great potential of our proposed framework for crowd counting in real-world scenarios.

Chapter 2

XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage

In this chapter, we propose a WiFi-video cross-modal person identification system. More specifically, given the WiFi CSI magnitude measurements of a pair of WiFi transceivers, obtained in an area where a person is walking, and the video footage of a person in another area, we propose a system that determines whether this given pair of video and WiFi measurements correspond to the same person or not. Unlike existing RF-based person identification systems, our system does not need prior wireless or video measurements of the person-of-interest for training purposes. It further does not need prior measurements in the operation environments.

The overall architecture of our proposed XModal-ID system and the various steps involved in the pipeline are shown in Fig. 2.1, and briefly described below:

- Given the video footage of a person, we construct a 3D mesh model of the person. We then propagate this mesh model over time and use Born approximation to simu-

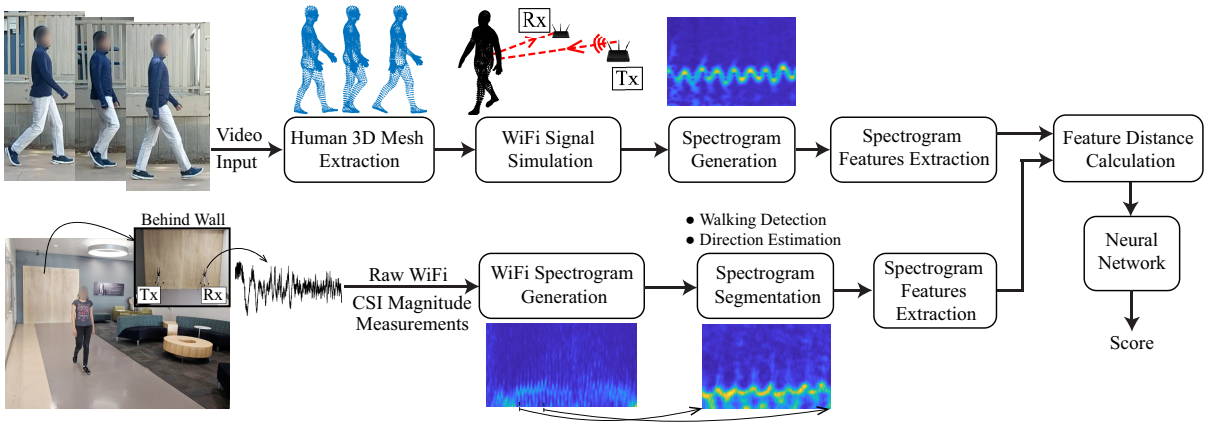


Figure 2.1: System architecture showing the various steps involved in the video and WiFi pipelines of XModal-ID. We refer readers to the color pdf for optimal viewing of the sample spectrograms.

late the corresponding received WiFi signal if the person was walking near a pair of WiFi transceivers. We then use the signal magnitude to generate the spectrogram of the signal, using Short-Time Fourier Transform (STFT). It is noteworthy that we do not need to know the track of the person or details of the operation area.

- In the operation area where a person is walking, a WiFi receiver (Rx) measures the CSI magnitude of the received signal in the transmission from a WiFi transmitter (Tx). We then generate the corresponding spectrogram from this CSI magnitude measurement (using a combination of STFT and Hermite functions) and segment it to obtain the parts most informative for identification, and further estimate the direction of motion.

- We then show how to extract key features from the spectrograms generated from both the WiFi and video data, and calculate the distance between them. The feature distances of a training set are fed into a small 1-layer neural network, which, after training, outputs a score indicating the similarity between any pair of real and simulated spectrograms, thus indicating if the person in a video is the same person in a WiFi area.

Figure 2.2: (Right) Three sample HMR algorithm output meshes for (left) different snapshots of a walking person.

2.1 Proposed XModal-ID System

In this section, we lay out the details of our proposed system, which is shown in Fig. 2.1. We first show how we can use a video footage of a walking person to generate a simulated wireless signal, which would have been measured if that person walked in a WiFi-covered area. Then, we show how to process the raw WiFi magnitude data measured in a real WiFi-covered area in which a person is walking. We mathematically model the wireless signals reflected from the person's body and apply time-frequency analysis techniques to generate a *spectrogram*, which captures the gait attributes of the person. We further focus on extracting the informative parts of the spectrogram as well as the direction of motion. We finally show how we can utilize the simulated wireless signal from the video footage to generate a corresponding spectrogram of the person based on the video. In Sec. 2.2, we then introduce a set of key features and show how we can use them to quantify the similarity between the two spectrograms to determine if they belong to the same person or two different people.

2.1.1 Video-to-WiFi Gait Modeling

In this section, we show how we can use a video footage of a walking person to generate a simulated WiFi signal, which would have been measured by a pair of WiFi

transceivers if this person walked in their vicinity. Note that we do not assume that the real WiFi transceivers are in the same area where the video footage was taken.

Given one video frame (snapshot) of a person, we first utilize the Human Mesh Recovery (HMR) algorithm of [62] to produce a dense 3D mesh, which contains a large number of 3D points describing the outer surface of the human body. Given a video clip of a person, we then construct a set of 3D points for each frame. The sequence of such sets then captures the gait of the person. Fig. 2.2 shows a few sample video snapshots with their corresponding 3D mesh models.

Denote by $\mathcal{M}(t) = \{\mathbf{p}_m(t) \in \mathbb{R}^3, m = 1, \ldots, M\}$ the set of generated 3D mesh points of the human body at time t.[1] In the real WiFi environment, a WiFi Tx is located at $\mathbf{p}_T \in \mathbb{R}^3$, and a WiFi Rx is located at $\mathbf{p}_R \in \mathbb{R}^3$, as shown in the bottom row of Fig. 2.1. In order to simulate the WiFi signal that would have been received if the person in the video was walking in the WiFi area, we utilize the Born approximation [64] to model the WiFi reflections off of the generated human mesh surface. More specifically, the simulated received WiFi signal at time t can be written as,

$$c_v(t) = \underbrace{\mathbf{g}(\mathbf{p}_T, \mathbf{p}_R)}_{\substack{\text{direct signal} \\ \text{from Tx to Rx}}} + \sum_{m \in \mathcal{M}'(t)} \underbrace{F_m K_m \mathbf{g}(\mathbf{p}_T, \mathbf{p}_m)\mathbf{g}(\mathbf{p}_m, \mathbf{p}_R)}_{\text{reflected signal from point } \mathbf{p}_m}, \tag{2.1}$$

where $\mathbf{g}(\mathbf{p}_1, \mathbf{p}_2)$ is the Green's function from point $\mathbf{p}_1$ to point $\mathbf{p}_2$ in $\mathbb{R}^3$, and is given by $\mathbf{g}(\mathbf{p}_1, \mathbf{p}_2) = \frac{\exp(\jmath \frac{2\pi}{\lambda} \|\mathbf{p}_1 - \mathbf{p}_2\|)}{4\pi \|\mathbf{p}_1 - \mathbf{p}_2\|}$, where $\|.\|$ is the Euclidean norm of the argument, and λ is the wavelength of the wireless signal. $\mathcal{M}'(t) \subset \mathcal{M}(t)$ is then the subset of all points in the human mesh that are visible to both the Tx and Rx, since only these points will reflect the signal to the Rx. We determine $\mathcal{M}'(t)$ by applying the Hidden Point Removal

[1] Note that the HMR method outputs 3D points in the pixel space. Transforming these points to real-world 3D coordinates only requires a one-time calibration of the camera upon fixation, using the coordinates of a few known points in the real world that are identified within the camera frame. See [63] for more details.

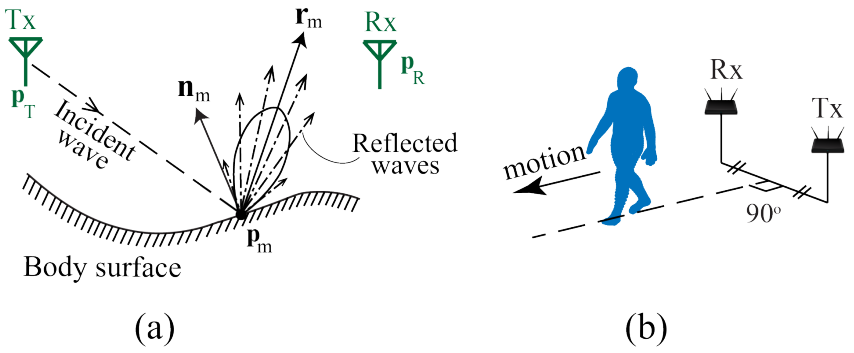


Figure 2.3: (a) Quasi-specular reflection model of the human body. An incident wave on a point $\mathbf{p}_m$ on the body is reflected to different directions with different amplitudes, with the strongest reflection being in the direction $\mathbf{r}_m$ determined by the normal to the body surface at $\mathbf{p}_m$. (b) Walking path of the generated human mesh to simulate the WiFi signal.

(HPR) algorithm [65] to $\mathcal{M}(t)$.

The strength of the signal reflected from point $\mathbf{p}_m$ is determined by two factors: the surface area and the orientation of the body part to which $\mathbf{p}_m$ belongs. For instance, the human torso has a higher reflectivity than the other body parts since it has a larger surface area. This factor is captured by the scale F_m. The orientation of the body part then determines the direction in which an incident signal would be reflected. A perfect reflector would reflect the incident wave at $\mathbf{p}_m$, only in the direction $\mathbf{r}_m = \frac{\mathbf{p}_m - \mathbf{p}_T}{\|\mathbf{p}_m - \mathbf{p}_T\|} - 2\frac{(\mathbf{p}_m - \mathbf{p}_T)^\intercal \mathbf{n}_m}{\|\mathbf{p}_m - \mathbf{p}_T\|}\mathbf{n}_m$, where $\mathbf{n}_m$ is the normal vector to the body at point $\mathbf{p}_m$ (see Fig. 2.3 (a)). However, the human body is best modeled as a *quasi-specular reflector* [66], which reflects the signal into many directions with different amplitudes, with the strongest in the $\mathbf{r}_m$ direction (as shown in Fig. 2.3 (a)). The amplitude of the reflection from $\mathbf{p}_m$ towards the Rx will then be inversely related to the angle between the vectors $\mathbf{p}_R - \mathbf{p}_m$ and $\mathbf{r}_m$. Based on our empirical studies, we capture this relation using a Gaussian mask $K_m = \exp\left(-\left(\cos^{-1}\frac{(\mathbf{p}_R - \mathbf{p}_m)^\intercal \mathbf{r}_m}{\|\mathbf{p}_R - \mathbf{p}_m\|}\right)^2 / 2\sigma_a^2\right)$.

We simulate the received wireless signal for the case where the person in the video is walking away from the link, on the line that is the perpendicular bisector of the Tx-Rx

link, as shown in Fig. 2.3 (b). We shall see in Sec. 2.1.3 why we do not need to know the real track of the person in the WiFi area and that simulating the receptions on only the aforementioned path will be sufficient for our XModal-ID system.

2.1.2 WiFi-Based Gait Modeling

In this section, consider the WiFi-covered area where a person is walking, as shown in Fig. 2.4. A WiFi Tx emits a wireless signal that reflects off of different parts of the human body and is received by a WiFi Rx. The complex baseband received signal $c(t)$ can be written as follows [67],

$$c(t) = \alpha_{\mathrm{s}} e^{j\mu_s} + \sum_m \alpha_m e^{j\left(\frac{2\pi}{\lambda}\psi v_m(t)t + \frac{2\pi}{\lambda} l_m\right)}, \tag{2.2}$$

where $\alpha_{\mathrm{s}} e^{j\mu_s}$ is the complex received signal including the impact of both the direct path and the static paths, α_m is the amplitude of the signal path reflected off of the m^{th} part of the body, l_m is the length of that path at time $t = 0$, $v_m(t)$ is the speed of the m^{th} body part at time t, and $\psi = \cos\phi_R + \cos\phi_T$ where ϕ_R and ϕ_T are as illustrated in Fig. 2.4.

Denoting by $s(t)$ the magnitude square of the baseband signal $c(t)$ and assuming that $|\alpha_{\mathrm{s}}| \gg |\alpha_m|$, $s(t)$ can be written as follows,

$$s(t) = P + \sum_m 2|\alpha_s \alpha_m| \cos\left(\frac{2\pi}{\lambda}\left(\psi v_m(t)t + l_m\right) - \mu_s\right), \tag{2.3}$$

where $P = |\alpha_{\mathrm{s}}|^2 + \sum_m |\alpha_m|^2$ is the DC component of $s(t)$. Note that ψ can be time-varying.

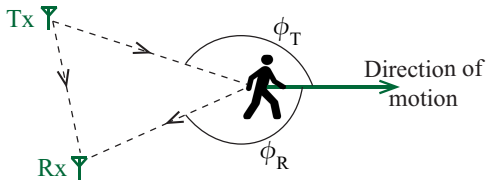


Figure 2.4: A pair of WiFi transceivers are used to identify the walking person.

2.1.3 Spectrogram Generation Based on Measured Wireless Signals

It can be seen from Eq. 2.3 that the signal $s(t)$ is the sum of multiple sinusoids whose frequencies are linearly related to the respective speeds of different body parts of the moving person. Hence, estimating the instantaneous frequency components of the signal $s(t)$ provides information about how the person walks. To this end, we utilize the Short-Time Fourier Transform (STFT), which is a commonly-used time-frequency analysis technique in the RF-based gait analysis literature. In STFT, a short moving window of length T_{win} is applied to $s(t)$ and the Fourier Transform is applied to each instance of the moving window to estimate the frequency components, resulting in a signal *spectrogram*. More specifically, we have,

$$STFT(t, f) = \left| \int_{t}^{t+T_{\text{win}}} s(\tau) e^{-j2\pi f \tau} d\tau \right|. \tag{2.4}$$

Fig. 2.5 (a) shows a sample STFT spectrogram of a walking person, which is generated from the received WiFi signal when a person walks away from a WiFi link, on a path perpendicular to it. A strong reflection (indicated by brighter colors) can be seen in the spectrogram at ∼25 Hz, which corresponds to a speed of 0.72 m/s. This is caused by the motion of the torso, the reflection of which is stronger due to its larger surface area. Weaker reflections (indicated by darker colors) of the faster body parts (e.g., legs) appear periodically at higher frequencies in the spectrogram.

While the STFT provides valuable information about the instantaneous speeds of different body parts, it has been shown in the literature that the corresponding time-frequency resolution trade-off can affect the quality of this information [68]. Multi-window Hermite Spectrograms (HS) were then proposed, in the Radar literature [68], to improve the concentration of STFT spectrograms. In a Hermite spectrogram, multiple Hermite functions are used as windows for the time-frequency analysis. More specifically,

$$HS(t,f) = \frac{1}{2\pi} \sum_{k=0}^{N_{\text{Hermite}}-1} b_k(t) \left| \int s(\tau)\chi_k(\tau - t)e^{-j2\pi f\tau} d\tau \right|^2, \tag{2.5}$$

where $\chi_k(t)$ is the k^{th} Hermite function, and $b_k(t)$ are weighting coefficients obtained by solving the system

$$\sum_{k=0}^{N_{\text{Hermite}}-1} b_k(t) \frac{\int |s(t+\tau)|^2 \chi_k^2(\tau)\tau^{n-1} d\tau}{\int |s(t+\tau)|^2 \chi_k^2(\tau) d\tau} = \begin{cases} 1, & \text{if } n = 1 \\ 0, & \text{if } n \in \{2, ..., N_{\text{Hermite}}\} \end{cases} \tag{2.6}$$

Fig. 2.5 (b) shows a Hermite spectrogram (with $N_{\text{Hermite}} = 3$ Hermite functions) generated from the same data as the STFT spectrogram in Fig. 2.5 (a). These two transformations, however, are utilized independently in different literature. In order to combine the desirable concentration properties of the HS and the ability of STFT to detect minute reflections from different body parts, we propose to generate the final WiFi spectrogram $S(t,f)$ by combining the two spectrograms as follows,

$$S(t,f) = STFT(t,f) + HS(t,f). \tag{2.7}$$

Essentially, $S(t,f)$ is a multi-window spectrogram that utilizes the rectangular window as well as the hermite function windows. We have observed that this combination considerably improves the visibility of the gait information in the Fourier domain. We then

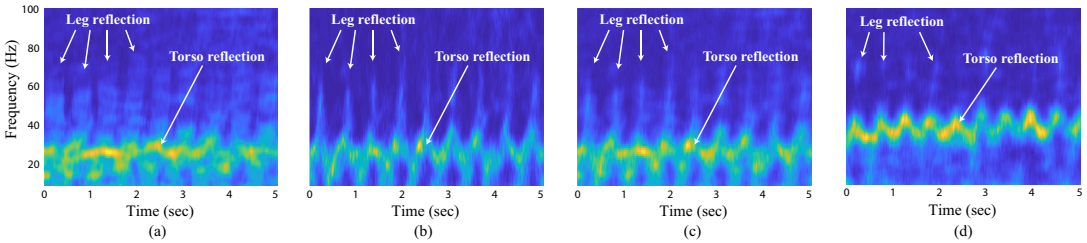


Figure 2.5: (a) Spectrogram based on Short-Time Fourier Transform (STFT) for a person walking away from the link, on a straight line perpendicular to the Tx-Rx line. (b) Spectrogram of the same data based on the Hermite method. (c) Combined Spectrogram $S(t, f)$ of STFT and the Hermite method. (d) Combined spectrogram $S(t, f)$ of another person walking on the same path. It can be seen from (c) and (d) that the combined spectrograms of different people are well differentiable, e.g., the torso speed, leg speed, and gait cycles are different. See the color pdf for optimal viewing of the spectrograms.

normalize the resulting spectrogram at each time instant, with respect to the sum of the values over all the frequencies at that time instant.

To visualize the impact of combining the spectrograms, consider the spectrograms shown in Fig. 2.5 (a) for STFT and Fig. 2.5 (b) for HS. As can be seen, the reflections of the person's limbs are clearly visible in the STFT, while the concentration of the torso reflection is clearer in the Hermite spectrogram. Consequently, the combined spectrogram in Fig. 2.5 (c) captures both these aspects of the gait. Additionally, Fig. 2.5 (d) shows the combined spectrogram of another subject walking on the same path, showing differentiable gait attributes.

Remark 2.1 *From Eq. 2.3, a detected reflection at a frequency f in the spectrogram is caused by a moving object with the speed $v = f\lambda/\psi$. Hence, static multipath due to the background environment appears at $f = 0$ and thus does not affect the gait motion information, which appears at non-zero frequencies.*

Remark 2.2 *We extract the gait information from the frequency of the reflected signal, and not from its power. Hence, as long as the power of the reflected signal is above the*

noise floor, the gait information can be extracted from the spectrogram. This is particularly attractive for through-wall settings, where the wall attenuates the signal power, but does not affect the gait motion information.

Spectrogram Segmentation:

As described in Eq. 2.3, two parameters determine the instantaneous frequencies of the different sinusoidal components of $s(t)$: the direction of motion (represented by ψ) and the instantaneous speeds of different body parts (v_m). In this section, we describe how we segment the spectrogram $S(t, f)$ and extract the part in which ψ can be considered constant. When the WiFi Tx and Rx are close to each other, as compared to the distance of the person to the link, a segment with an approximately constant ψ is obtained whenever the person walks on a straight line towards or away from the midpoint of the Tx-Rx line. In such a segmented spectrogram, the frequency information mainly contains the gait attributes of the person, since it depends only on the speeds of the different body parts (v_m). **As such, it can be very informative for person identification, without requiring the knowledge of the track of the person.** We henceforth refer to such a segment as a *constant-ψ segment*, and utilize it for our XModal-ID system. Note that we do not need the whole track to be on a straight line towards/away from the midpoint of the link. The person can take any track, and as long as there is even a small part of the track (e.g., 3 sec or longer) that satisfies this condition, then the proposed approach can be utilized.

In order to extract a constant-ψ segment from the spectrogram, we search for a segment (with a minimum width of $T_{\min}$) that satisfies two conditions. First, the spread of the energy distribution across frequency inside the segment, $V(t) = \int f^2 S(t,f)df - \left(\int fS(t,f)df\right)^2$, should remain below a certain threshold V_{th}, since a higher value of $V(t)$ indicates that the spectrogram is close to being flat at time t, which implies that there

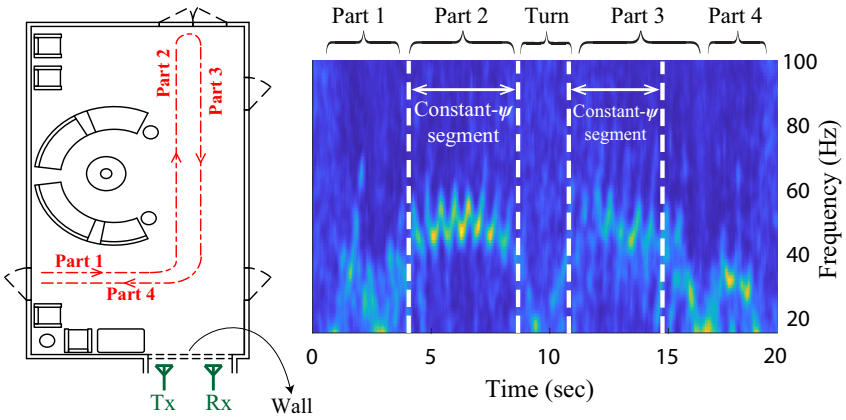


Figure 2.6: Working example of the spectrogram segmentation algorithm. (Left) A floor plan with a 4-part path where a person walks (experiment area of Fig. 2.9 (g)), with WiFi Tx-Rx placed behind a wall. (Right) The spectrogram of the measured WiFi data, showing different parts of the walk. The dashed lines show two extracted constant-ψ segments.

is no walking detected within this segment. Secondly, the variations of the average torso speed within this segment should remain below a certain threshold v_{th}. Since the average torso speed of a walking person is constant in a small time window, a varying average torso speed in the spectrogram is due to a varying ψ. The average torso speed can be calculated from the spectrogram, as we shall see in Sec. 2.2. When a segment satisfies the aforementioned conditions, it is declared as a constant-ψ segment.[2]

Next, we consider what would be a good value for T_{min} (the minimum acceptable width of the segment). A small T_{min} would result in many false positives, in which ψ could be falsely considered constant simply because the segment was too short. On the other hand, a large T_{min} would require the person to walk for a long time in order to be identified. We observe that using $T_{\text{min}} = 3$ sec is a good trade-off, which provides a sufficient number of gait cycles (for a casual walk) for extracting meaningful gait features.

Fig. 2.6 shows an example of the spectrogram segmentation algorithm for the walking experiment depicted on the left, where the constant-ψ portion corresponds to parts 2 and

[2]Note that there can be multiple constant-ψ segments in one spectrogram, depending on the track of the person, which we assume unknown.

3 of the track. The figure on the right shows the un-segmented spectrogram $S(t, f)$ of the entire walking experiment, as well as the constant-ψ segments detected by our algorithm.

Walking Direction Estimation:

We have observed that the segments of the spectrogram corresponding to a person walking away from the link have clearer gait patterns than those corresponding to walking towards the link (for instance, compare parts 2 and 3 in Fig. 2.6). Similar observations have been made in [69]. Therefore, we propose to utilize only the spectrogram segments corresponding to when the person is walking away from the link in our subsequent processing pipeline. Let $S_w(t, f)$ denote a constant-ψ spectrogram segment detected by the spectrogram segmentation algorithm. The information about the direction of motion, i.e., whether the person is walking towards or away from the link, is theoretically contained in the sign of ψ. However, this information cannot be extracted from $S_w(t, f)$ since both a positive and a negative ψ would result in the same spectrogram, given that we only use signal magnitude measurements. In this section, we then propose a new method that can estimate the walking direction.

Despite the absence of the information about the sign of ψ in $S_w(t, f)$, we can still determine the walking direction by exploiting the fact that a WiFi signal spans frequencies in the band $[f_c - B/2, f_c + B/2]$, where f_c is the carrier frequency and B is the WiFi bandwidth, as we shall see next. Based on Eq. 2.3, the magnitude squared WiFi signal, $s(t; \rho)$, measured in a very short time on a frequency range $f_c + \rho$, for $\rho \in [-B/2, B/2]$,

can be written as,

$$\begin{aligned}\bar{s}(t;\rho) &= s(t;\rho) - P \\ &= \sum_m 2|\alpha_s\alpha_m| \cos\left(\frac{2\pi}{c}(f_c+\rho)(v_m\psi t + l_m) - \mu_s\right) \\ &\approx \sum_m 2|\alpha_s\alpha_m| \cos\left(\frac{2\pi}{c}\left(f_c v_m \psi t + f_c l_m + \rho l_m\right) - \mu_s\right), \end{aligned} \tag{2.8}$$

where c is the speed of light. The approximation in the last line of Eq. 2.8 relies on the fact that, in a very short time window, the product ρt is negligible as compared to the other terms in the cosine argument. By taking the Fourier Transform of $\bar{s}(t;\rho)$ along the t dimension and the inverse Fourier Transform along the ρ dimension, we get,

$$\begin{aligned} Z(f;\zeta) &= \left| \iint \bar{s}(t;\rho) e^{-j2\pi f t} e^{j2\pi\zeta\rho} dt d\rho \right| \\ &= \sum_m |\alpha_s\alpha_m| \left(\delta\left(f - \frac{v_m\psi}{\lambda_c}, \zeta - \frac{l_m}{c}\right) + \delta\left(f + \frac{v_m\psi}{\lambda_c}, \zeta + \frac{l_m}{c}\right)\right), \end{aligned} \tag{2.9}$$

where $\delta(.,.)$ is the 2D Dirac Delta function, and $\lambda_c = c/f_c$. By examining Eq. 2.9, we can see that for a positive ψ (a person moving towards the link), the components of Z lie in the first and third quadrants of the (f,ζ) space, while, for a negative ψ (a person moving away from the link), the components of Z lie in the second and fourth quadrants of the (f,ζ) space. Therefore, we can determine the walking direction of the person according to the following decision rule,

$$\frac{\int_0^\infty \int_0^\infty |Z(f;\zeta)|^2 df d\zeta}{\int_0^\infty \int_{-\infty}^0 |Z(f;\zeta)|^2 df d\zeta} \underset{\text{away}}{\overset{\text{towards}}{\gtrless}} 1\ . \tag{2.10}$$

Fig. 2.7 shows an example of the direction estimation output $Z(f;\zeta)$ when a person is walking towards the link (left), and away from the link (right). It can be clearly seen that the energy of $Z(f;\zeta)$ is concentrated in the first and third quadrants for the former

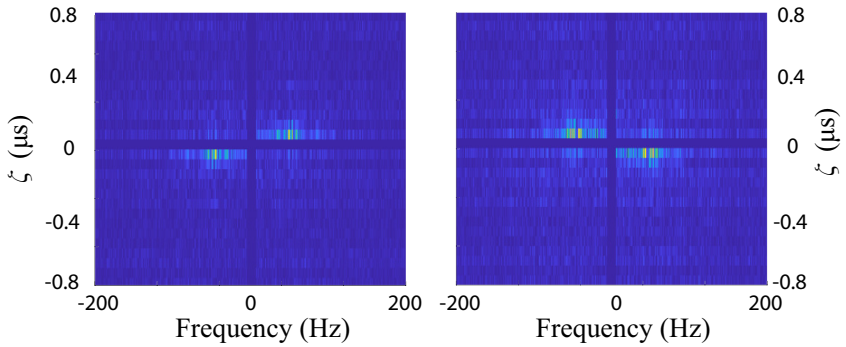


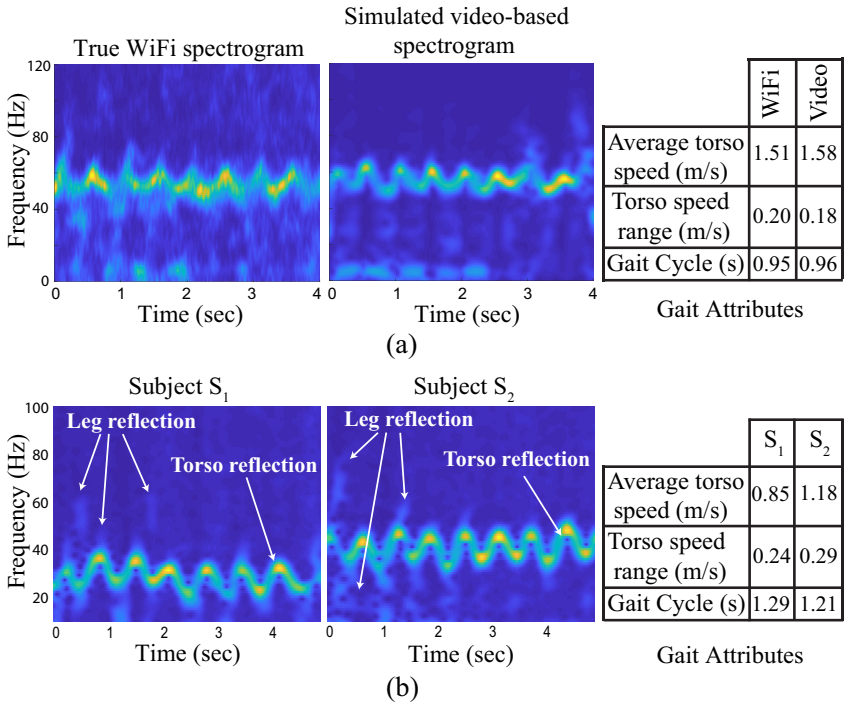
Figure 2.7: Plots of $Z(f, \zeta)$ when a person is walking (left) towards the link, and (right) away from the link. Energy distribution of Z over the four quadrants indicates the motion direction. See the color pdf for better viewing.

case, and in the second and fourth quadrants for the latter case.

2.1.4 Video-Based Spectrogram Generation

In Sec. 2.1.1, we proposed a way of simulating the wireless signal based on the video footage of the person, using the HMR algorithm, HPR algorithm, and Born approximation. Since we are only interested in the constant-ψ parts of the spectrogram of the real WiFi measurement, as discussed in Sec. 2.1.3, we only need to simulate the corresponding wireless signal of Eq. 2.1 when a person walks away from the link, on the line that is the perpendicular bisector of the link (see Fig. 2.3 (b)), while being far enough from the link, as compared to the distance between the Tx and Rx (e.g., at least 2 m when the Tx and Rx are 1.5 m apart). As such, no knowledge of the track or operation area is needed. After the calculation of $c_v(t)$ using Eq. 2.1 on the aforementioned path, a spectrogram of the simulated WiFi signal $S_v(t, f)$ is generated from $|c_v(t)|^2$ via STFT, using the procedure described by Eq. 2.4.

To validate our framework of video-based spectrogram generation, we conduct a preliminary experiment where a subject walks away on a path that perpendicularly bisects the WiFi link, while simultaneously being videotaped. Fig. 2.8 (a) shows the spectro-



	WiFi	Video
Average torso speed (m/s)	1.51	1.58
Torso speed range (m/s)	0.20	0.18
Gait Cycle (s)	0.95	0.96

Gait Attributes

	S_1	S_2
Average torso speed (m/s)	0.85	1.18
Torso speed range (m/s)	0.24	0.29
Gait Cycle (s)	1.29	1.21

Gait Attributes

Figure 2.8: (a) Spectrograms of real WiFi data and of the video-based simulated one for the same person, showing similar gait attributes. (b) Video-based simulated spectrograms of two different people, showing their distinct gait attributes.

gram of the real WiFi data as well as the video-based simulated spectrogram. The figure further shows some gait attributes extracted from both spectrograms. It can be seen that our video-based spectrogram closely resembles the real WiFi data spectrogram, demonstrating the accuracy of our proposed framework. Additionally, Fig. 2.8 (b) shows two sample video-based spectrograms of two different subjects. It can be seen that the gait attributes of the two subjects are well differentiable in the two spectrograms.

Next, we show how to measure the similarity between the video-based simulated spectrogram and the spectrogram obtained from the real measured WiFi data, in order to identify whether the video and the WiFi data correspond to the same person or two different people.

2.2 Feature Extraction and Similarity Prediction

So far, we have described our approach to extract spectrograms from both WiFi data and video data. In order to determine whether a WiFi sample and a video sample belong to the same person, we extract several features from the corresponding spectrograms. We then compute a set of distances between the WiFi-based and video-based features. Given these feature distances, we train a simple 1-layer neural network to properly combine the distances and determine if a pair of WiFi and video samples belong to the same person. After training, the network not only provides this binary prediction, but also provides a score indicating the similarity between the WiFi and video samples. Once the network is trained, we use it on unseen WiFi and video data. In other words, none of the test data and locations is used for training.

2.2.1 Spectrogram Features

We have identified 12 features that are key for capturing the main characteristics of a person's gait. We compute each feature on both the WiFi and video spectrograms, and use a distance metric to measure the difference between the two spectrograms with respect to each feature. More specifically, we look at the frequency and time dimensions of the spectrogram, which carry different types of gait signatures that can be used for identification, as we describe next.

The frequency dimension carries information about the speeds of different body parts. We use the following frequency-related features:

• **Frequency distribution (FD):** This feature is obtained by averaging the spectrogram over time. This feature captures the distribution of frequency components, or equivalently, the speeds of different body parts, during the person's walk.

• **Frequency distribution in** 4 **gait phases (FD4):** Similar to the previous feature,

we calculate the time-average of the spectrogram for each of the 4 phases of the gait cycle, resulting in 4 corresponding feature vectors [1].

• **Average torso speed:** We calculate the average of the torso speed curve, which can be extracted from the spectrogram using the method in [1].

• **Average of the range of torso speed:** After extracting the torso speed curve, we calculate the range of the torso speed variation in one gait cycle. This range is then averaged over all the gait cycles in the spectrogram.

The time dimension carries temporal information (e.g., periodicity patterns) about a person's gait. We capture the temporal signatures using the following features:

• **Autocorrelation (AC):** Given a spectrogram, we compute the autocorrelation across time (with a maximum lag of 2 sec) for each frequency bin, resulting in an autocorrelation matrix. We then compute a weighted sum of this matrix over the frequency dimension based on the energy distribution over the frequencies. This feature carries information on the gait cycle and the periodicity of the walk.

• **FFT of spectrogram over time:** Similar to the method in [69], we calculate the Fast Fourier Transform (FFT) of the spectrogram over time for each frequency bin. We then compute a weighted sum over the frequency dimension based on the energy distribution over the frequencies.

• **Histogram of autocorrelation gradient:** This is the histogram of the gradient of the AC feature vector.

• **Histogram of torso speed gradient:** We calculate the histogram of the torso speed gradient, which carries information on how the torso speed changes with time.

• **Stride length:** This is obtained by multiplying the average torso speed by the gait cycle length, which can be extracted from the torso speed curve.

Given the 12 features of a WiFi spectrogram and the 12 features of a video-based simulated spectrogram, we compute the distance between each corresponding feature in

WiFi and video. This results in a vector of 12 feature distances. More specifically, for the frequency distribution (FD), we use the Kullback-Leibler Divergence (KLD) as the distance metric. For the frequency distributions over 4 gait phases (FD4), for each gait phase, we first align the WiFi-based and video-based features by offsetting their respective average torso speeds. We then use KLD as the distance metric between the two aligned features. The alignment removes the effect of area-dependent average speeds (see Sec. 2.5) and places more focus on the relative speeds of body parts. For autocorrelation (AC), we use the cosine similarity. For all the other features, we use the Euclidean distance.

2.2.2 Similarity Prediction

Given a pair of WiFi and video data samples, we compute a set of 12 distances as described previously. We then utilize a simple neural network to combine these distances into a final decision on whether this WiFi-video pair belongs to the same person. We train the network on WiFi/video data and locations disjoint from the test subjects and areas (more details in Sec. 2.3). During training, these 12 distances are fed into the neural network, which has 1 hidden layer with 30 units, along with a binary label indicating whether these two samples belong to the same person. After training, the network can provide a binary decision on a given pair and a confidence score indicating the similarity between the pair.

2.3 Experimental Setup and Data Collection

In this section, we describe the experimental setup for collecting data (both wireless and video) and validating our proposed methodology. We then show how we construct the training set for training a small neural network described in Sec. 2.2.2, and the test set for evaluating our proposed system.

2.3.1 Experiment Subjects

In order to collect WiFi and video data, we have recruited a total of 18 subjects. We divide them into two disjoint sets of 10 and 8 subjects, for training and test, respectively. As a result, the test set consists of the walking data of people that have never been seen during training, which allows us to evaluate the proposed system's ability to generalize to new people. In the training set, the 10 subjects (referred to as the training subjects) consist of 9 males and 1 female, with heights ranging from 163 cm to 186 cm. In the test set, the 8 subjects (referred to as the test subjects) consist of 6 males and 2 females, with heights ranging from 160 cm to 186 cm. The speeds of the test subjects have a mean of 1.43 m/s and a standard deviation of 0.26 m/s, while their gait cycles have a mean of 1.06 sec and a standard deviation of 0.17 sec.

2.3.2 WiFi Data Collection

In this part, we describe the experiments where we use a pair of WiFi transceivers to collect the WiFi data of the subjects.

Experiment Platform and Data Processing:

For the WiFi data collection process, we use two laptops equipped with Intel 5300 WLAN Network Interface Cards (NICs). We mount $N_{\text{Tx}} = 2$ omni-directional antennas to a tripod of height 85 cm, and connect them to two antenna ports on the Tx laptop, which transmits WiFi packets on WiFi channel 36 with a carrier frequency of 5.18 GHz. Similarly, we mount $N_{\text{Rx}} = 2$ receiving antennas to a tripod of the same height, located 1.5 m away from the Tx antennas, and connect them to two antenna ports on the Rx laptop, which logs the CSI information on 30 subcarriers with a rate of 2,000 packets/sec. The data is then processed offline to extract the CSI information using Csitool [70]. The

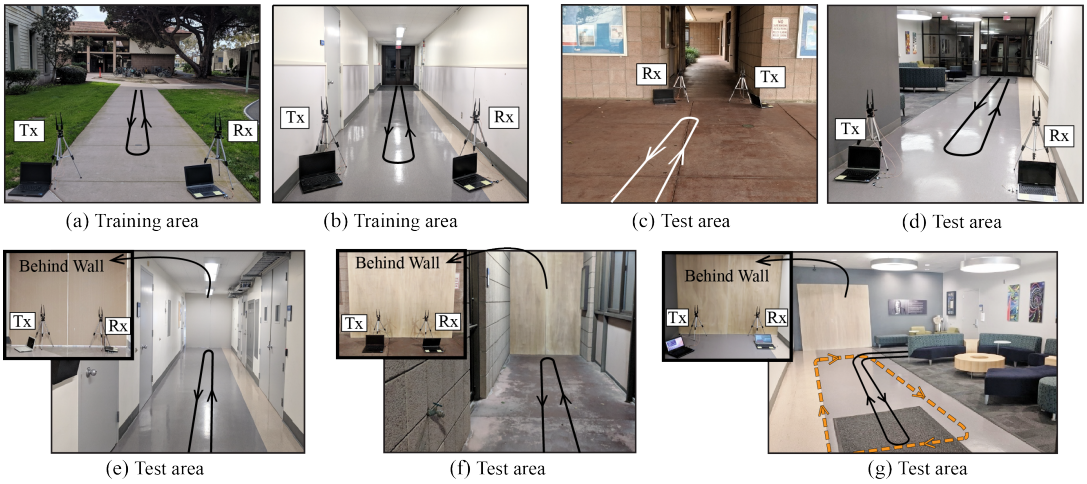
setup results in a total of $N_{\text{Tx}} \times N_{\text{Rx}} \times 30 = 120$ streams of data which we process using the method in [1]. More specifically, we denoise the data using Principal Component Analysis (PCA). We first generate spectrograms of the first 15 PCA components of the measured signal, using time windows of $T_{\text{win}} = 0.4$ sec, with a shift of 4 ms. We then average these 15 spectrograms to obtain the final spectrogram. The frequency axis ranges from 15 Hz to 125 Hz (which translates to speeds of 0.4 m/s to 3.6 m/s). For the spectrogram segmentation algorithm, we set $V_{\text{th}} = 0.8$ for indoor areas and 0.88 for outdoor areas. These values were determined by using the experimental data of 3 training subjects. We also set the allowable change in average speed to $v_{\text{th}} = 0.3$ m/s.

Experiment Scenarios:

In the WiFi experiments, we use three different settings for collecting the WiFi CSI data, as described below and shown in Fig. 2.9:

• **Line-of-Sight Straight-Path (LOS-SP):** In this setting, a WiFi link is deployed in the environment where the person is walking, with a direct view of the person. In each experiment, the subject walks from a starting point that is at least 8 m from the link and towards the link. The subject turns around when he/she is ∼1 m away from the link and then walks back to the starting point. This setting captures how people typically walk in a hallway or a pathway environment. The corresponding areas are shown in Fig. 2.9 (a) - (d). Areas of Fig. 2.9 (a) and (b) are only used for training while areas of Fig. 2.9 (c) and (d) are only used for testing.

• **Through-Wall Straight-Path (TW-SP):** In this setting, the subjects walk on a path similar to the LOS-SP setting. However, in this case, the WiFi Tx and Rx are placed behind a wall, without any view of the walking subject. We use plywood and drywall for the through-wall experiments, which are used for the walls of ∼90% of residential and small commercial buildings in the U.S. [71], hence, showing the applicability of our



(a) Training area (b) Training area (c) Test area (d) Test area

(e) Test area (f) Test area (g) Test area

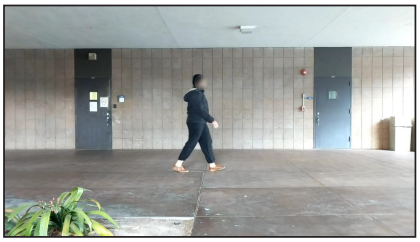
Figure 2.9: (a) – (b) WiFi training areas, Line-of-Sight Straight-Path (LOS-SP) setting: we collect WiFi CSI data of the training subjects in these two areas. (c) – (d) WiFi test areas, Line-of-Sight Straight-Path (LOS-SP) setting. (e) – (f) WiFi test areas, Through-Wall Straight-Path (TW-SP) setting. (g) WiFi test area, Through-Wall Complex-Path (TW-CP) setting, with two complex routes indicative of how people generally walk in this lounge area.

proposed system to typical through-wall environments. Our two TW-SP areas are shown in Fig. 2.9 (e) and Fig. 2.9 (f). TW-SP areas are only used for testing.

• **Through-Wall Complex-Path (TW-CP):** In this setting, the WiFi Tx and Rx are placed behind a wall. Unlike the previous straight-path settings, the subjects walk on more general and complex paths. As shown in Fig. 2.9 (g), in the TW-CP experiments, the subjects are asked to walk on two different complex paths that are representative of how people would typically walk in a lounge environment. TW-CP area and complex paths of Fig. 2.9 (g) are only used for testing.

Experiment Areas (see Fig. 2.9):

We use the walking data of the 10 training subjects in 2 LOS-SP areas (Fig. 2.9 (a) and (b)) for training the neural network. The walking data of the remaining 8 test subjects in the remaining 5 areas, 3 through-wall (2 TW-SP and 1 TW-CP) and 2 line-of-sight, is then used for testing. The training and test areas all vary in size and geometry.



(a) Video training area

(b) Video test areas

Figure 2.10: Sample snapshots for videos in (a) the training video location, and (b) the two test video locations.

In order to create more statistics and avoid biasing the results to a particularly favorable or unfavorable data point, each test subject walks back and forth in each area twice. Each such data instance (one back-and-forth) is then treated independently in the data pool.

2.3.3 Video Data Collection

In order to train and test XModal-ID, we collect the video data of the 18 subjects walking in an area. For training, we have collected videos of the 10 training subjects walking in one area, shown in Fig. 2.10 (a). For testing, we have collected videos of the 8 test subjects walking in two different areas shown in Fig. 2.10 (b). The video data collection areas are completely disjoint from the WiFi experiment locations. In each video area, a subject walks back and forth on a 7-m straight path and a side-view video (with a frame rate of 60 fps) is recorded. The videos are then manually clipped such that each resulting video clip contains a subject walking on a straight path in one

direction. Overall, each video clip has an average duration of 4.7 sec. Each video clip is then treated independently in the data pool. We collected a total of 100 such video clips of the training subjects and 96 clips of the test subjects (by having them repeat the back and forth path a number of times).

We process the frames of each video clip using the algorithm described in Sec. 2.1.1. The HMR algorithm outputs a total of 2,300 mesh points on the human body for each frame. The number of mesh point sets (frames) is then upsampled to have a frame rate of 250 fps. Based on the surface area values mentioned in [72], we approximate the reflectivity of the torso points to be 3 times the reflectivity of other body parts (which are all taken to have the same reflectivity). For the quasi-specular reflection beam, we set $\sigma_a^2 = 40$ based on the data of 3 training subjects.

To obtain the final video-based features of a walking person, we average the 12 features (described in Sec. 2.2) over 4 randomly-selected video-based spectrograms of that person (i.e., over 4 video clips of that person). Such averaging is feasible in practice as these 4 spectrograms can be generated from chunks of a longer video or from a few short video clips of the same person. In this chapter, the 4 spectrograms amount to a total video duration of 18.8 sec on average.

2.3.4 Training and Test Sets

Given the collected WiFi and video data, we construct a training set and a test set. For both sets, we first generate the spectrograms for the WiFi data samples and the video clips as described in Sec. 2.1. After the spectrogram generation, each training or test instance consists of a WiFi data sample and a video data sample (drawn from the corresponding training or test pools), a distance vector between their corresponding features, and a label indicating whether they belong to the same subject. A positive

label indicates that the pair belongs to the same person and a negative label denotes otherwise.

The training set is based on the 10 training subjects walking in the 2 WiFi training areas in the LOS-SP setting (Fig. 2.9 (a) and (b)) and in the 1 video training area (Fig. 2.10 (a)). The training set consists of a total of 7,280 pairs of WiFi-video instances. As we have a different number of pairs with positive and negative labels, we utilize oversampling [73] to obtain a balanced training set. The neural network discussed in Sec. 2.2.2 is then implemented in PyTorch [74].

The test set is based on the 8 test subjects' data in the 5 WiFi test areas (Fig. 2.9 (c) - (g)) and the 2 video test areas (Fig. 2.10 (b)). The test set includes all the 3 scenarios: LOS-SP (Fig. 2.9 (c) and (d)), TW-SP (Fig. 2.9 (e) and (f)), and TW-CP (Fig. 2.9 (g)) in the WiFi experiments. In the test set, each WiFi sample is paired with several randomly-selected video samples. Overall, we have a total of 2,256 instances (i.e., pairs of WiFi and video data samples) in the test set, with 768 pairs in the LOS-SP setting, 744 pairs in the TW-SP setting, and 744 pairs in the TW-CP setting. In addition to binary classification (i.e., does a WiFi-video pair belong to the same person or not?), we also test the ranking accuracy of our proposed system (see Sec. 2.4.1). In each ranking test, a WiFi sample serves as a query and 8 video samples serve as the candidates, with one of them containing the same subject as in the WiFi sample. We have a total of 282 such ranking tests in the test set, with 96 in the LOS-SP setting, 93 in the TW-SP setting, and 93 in the TW-CP setting.

2.4 System Evaluation

In this section, we present extensive experimental evaluations of our proposed system in various practical settings using a large test set. Unlike existing studies on WiFi-based

person identification, our test set only contains subjects and areas that have never been seen during the training process.

2.4.1 Evaluation Criteria

We use the following two evaluation criteria, which are both relevant in different applications:

1. Binary classification accuracy: In this setting, we evaluate our proposed system by using pairs of WiFi and video samples. Given a pair of WiFi and video data samples, the system predicts whether they belong to the same person or not. The resulting binary classification accuracy is used as the evaluation metric. As we have different numbers of test instances with positive (same-person) labels and negative (different-people) labels, we report the balanced classification accuracy, i.e., the average of the respective accuracies over the same-person and different-people pairs.

2. Ranking Accuracy: In each ranking test, the system is given a WiFi sample of a test subject and the video samples of several subjects from the test set. Among these candidate video samples, only one of them belongs to the person corresponding to the queried WiFi sample, to which we refer as the correct video sample. The system then ranks the video samples based on their similarity to the WiFi sample. We report the top-1, top-2, and top-3 ranking accuracies in this setting, where the top-k accuracy is defined as the percentage of cases where the correct video sample is ranked among the top k positions of all the video samples in a test.

Remark 2.3 *Note that if the number of subjects in the ranking test is 2, the system determines which one of the two video samples belongs to the person in the queried WiFi sample. This is different from the binary classification task, which determines whether a video sample and a WiFi sample belong to the same person or not.*

Area	Binary class. accuracy	Ranking accuracy		
		Top-1	Top-2	Top-3
Line-of-Sight Straight-Path setting				
Area of Fig. 2.9 (c)	90%	87%	96%	98%
Area of Fig. 2.9 (d)	86%	70%	83%	95%
Average	88%	78%	90%	96%
Through-Wall Straight-Path setting				
Area of Fig. 2.9 (e)	83%	74%	90%	97%
Area of Fig. 2.9 (f)	89%	82%	96%	100%
Average	86%	78%	93%	98%
Through-Wall Complex-Path setting				
Area of Fig. 2.9 (g)	82%	69%	86%	96%
Overall average	85%	75%	90%	97%

Table 2.1: The binary classification accuracy and top-1 to top-3 ranking accuracies of XModal-ID on the test set, in three different settings. The last row shows the average performance over all the areas/settings.

2.4.2 Performance Evaluation

In this section, we evaluate our proposed system on our extensive test set, which only has experimental areas and subjects that are not seen during the training phase. We further extensively test our system in through-wall scenarios and with complex paths. See Sec. 2.3.4 for the details of the test set. It is noteworthy to re-emphasize that our system does not need to know the track of the subject, or the details of the test area, as we discussed in Sec. 2.1.3. Furthermore, all the test videos are from areas of Fig. 2.10 (b) (disjoint from the WiFi areas), as discussed earlier. Table. 2.1 summarizes all the results that we shall discuss in this section.

Evaluation of Line-of-Sight Scenarios:

We first evaluate XModal-ID in the Line-of-Sight Straight-Path (LOS-SP) setting, consisting of 2 WiFi areas (Fig. 2.9 (c) and (d)). In this case, XModal-ID achieves a binary classification accuracy of 90% in the area of Fig. 2.9 (c) and 86% in the area of Fig. 2.9 (d), resulting in an overall average binary classification accuracy of 88%. In other words, given a pair of WiFi and video samples, both generated from subjects

and environments not seen during training, our system has an 88% chance of correctly predicting whether these two samples correspond to the same person or not, in these two areas.

Next, we look at the ranking performance. In the LOS-SP setting, given a queried WiFi sample and 8 candidate video samples of the 8 test subjects, XModal-ID has a success rate of 78% of assigning the highest rank to the correct video sample, in these two areas. Note that a random selection would only result in a success rate of 12.5%. Moreover, in this setting, XModal-ID has top-2 and top-3 accuracies of 90% and 96%, respectively. The ranking accuracy per area is shown in Table 2.1.

Evaluation of Through-Wall Scenarios:

Next, we consider the Through-Wall Straight-Path (TW-SP) WiFi areas (Fig. 2.9 (e) and (f)), where the WiFi link is placed behind a wall and does not have any view of the subjects. XModal-ID achieves a binary classification accuracy of 83% in the area of Fig. 2.9 (e) and 89% in the area of Fig. 2.9 (f), amounting to an overall average accuracy of 86%. In terms of ranking, XModal-ID achieves top-1, top-2, and top-3 accuracies of 78%, 93%, and 98%, over both areas. In particular, when XModal-ID is deployed in the area of Fig. 2.9 (f), it includes the correct video sample among the top 3 all the time.

In the Through-Wall Complex-Path (TW-CP) area, shown in Fig. 2.9 (g), the WiFi link is placed behind a wall and each test subject walks on two sample complex paths (each path treated as a separate experiment). These two paths represent how people would typically walk in this lounge area. In this setting, XModal-ID achieves a binary classification accuracy of 82%. For the case of ranking, our system obtains top-1, top-2, and top-3 accuracies of 69%, 86%, and 96%, respectively, in this area. It is noteworthy that in this TW-CP setting, which showcases challenging real-world application scenarios, the system has a very high probability (0.96) of including the correct video sample within

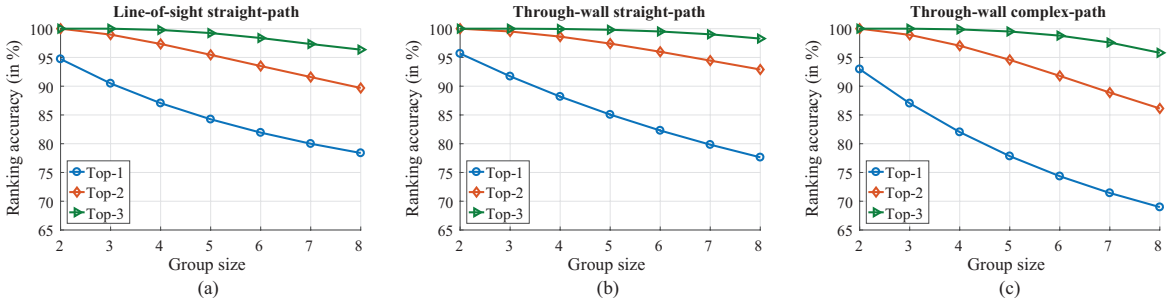


Figure 2.11: Top-1 to top-3 ranking accuracies when group size varies from 2 to 8, in (a) Line-of-Sight Straight-Path (LOS-SP) areas, (b) Through-Wall Straight-Path (TW-SP) areas, and (c) Through-Wall Complex-Path (TW-CP) area.

the top 3 ranks.

Overall, XModal-ID achieves a binary classification accuracy of 85%, and top-1, top-2, and top-3 ranking accuracies of 75%, 90%, and 97%, over all 5 areas/scenarios. These results demonstrate that XModal-ID has a robust performance, even when the transceivers are placed behind a wall, without any prior knowledge or view of the person/area, and when the subjects walk on unknown and complex paths.

2.4.3 Evaluation with Different Group Sizes

In the previous part, we showed the performance of our proposed system on the full test set consisting of 8 subjects. While the binary classification accuracy is independent of the number of subjects, ranking accuracy is a function of the number of subjects. In this section, we then study the performance of XModal-ID by varying the number of subjects in the test set, to which we refer as the group size.

Fig. 2.11 (a) shows the top-1, top-2, and top-3 ranking accuracies when the group size is varied from 2 to 8, in the LOS-SP setting. For each group size that is smaller than 8, the accuracies are averaged over all the possible subsets of subjects for that group size. As can be seen, as we reduce the group size, the ranking accuracies increase, since, with a smaller group size, it is less likely to have two subjects with similar gaits. When the

group size is less than 8, the top-1 ranking accuracy is always greater than 80%.

Fig. 2.11 (b) and (c) further show the ranking accuracies in the through-wall straight-path and complex-path settings, respectively, as a function of the group size. As can be seen, the accuracies increase as the group size decreases. Notably, when the group size is less than 8, the top-3 accuracy in these two through-wall settings is very close to 100%.

Overall, these evaluation results show that XModal-ID can successfully perform cross-modal person identification even when the test subjects and areas have never been seen before. The test set areas represent a wide variety of real-life scenarios, including through-wall scenarios and cases where the person walks on a complex path (rather than a straight one). Our system does not even need to know the track of the subjects. Overall, our results demonstrate the efficacy of XModal-ID in various real-world scenarios.

2.5 Discussions

In this section, we discuss a few key aspects of XModal-ID, as well as its limitations and future extensions.

Environment-Dependent Average Speeds: Environmental factors can sometimes affect people's average walking speed [75]. For instance, we noticed that people tend to walk slightly faster in outdoor/open areas, as compared to indoor/closed areas. All existing works on WiFi-based gait identification train and test in the same area, where the subjects mostly maintain the same walking speeds. On the other hand, in XModal-ID, in addition to the overall average speed, we also utilize spectrogram features that are independent of the average speed and only depend on the distribution of the relative speeds of body parts (see Sec. 2.2). Hence, XModal-ID can tolerate small changes in the average speeds of the subjects.

Tracks with Varying ψ: XModal-ID does not assume any knowledge of the track of the

person. Instead, it uses the spectrogram segmentation algorithm in Sec. 2.1.3 to extract the part of the person's track where ψ is approximately constant. The constant-ψ parts correspond to parts of track where the subject walks on a straight path towards/away from the midpoint of the Tx-Rx line, for the case where Tx and Rx are close enough to each other (see Sec. 2.1.3). Since this is a very general condition, most natural tracks will at least have small parts that would satisfy this condition. In fact, XModal-ID only needs a very small part of the track, e.g., 3 sec, to satisfy this condition, as discussed earlier. In the rare case that no part of the track satisfies this condition, the varying ψ can be estimated by existing WiFi-based tracking approaches and XModal-ID can be extended to accommodate the varying ψ.

Applicability to Intruder Detection: XModal-ID can also determine whether a WiFi sample belongs to a new user whose video is not available. It can compare this WiFi sample with each of the available video samples, using the binary classification criterion, and declare an unseen user if the WiFi sample does not match any of the videos. This setting can be relevant in applications such as intruder detection.

Processing Time: A typical duration of a WiFi data sample in our experiments is 25 sec. On a 3.40 GHz Intel Core i7 PC, XModal-ID takes an average of $\sim$19.8 sec to fully process such WiFi data. For videos, XModal-ID takes $\sim$132.5 sec to fully process a video clip of 4.7 sec (average duration) in order to generate a final feature vector. In particular, $\sim$112.8 sec are dedicated to generating the human mesh model, using the publicly available codes of [76, 62] on an NVidia GTX 1070 GPU, while the remaining steps (e.g., WiFi signal simulation) take $\sim$19.7 sec on a 3.40 GHz Intel Core i7 PC.

Chapter 3

Multiple People Identification Through Walls Using Off-The-Shelf WiFi

In this chapter, we extend the person identification ideas of Chapter 2 to perform through-wall person identification when multiple people are walking simultaneously in an area. Similar to the previous chapter, our system does not require any prior training measurement of the people to be identified. The overall system architecture is outlined in the flowchart of Fig. 3.1, and summarized below:

- Given the CSI magnitude of a WiFi signal measured at a small number of receivers when multiple people are walking simultaneously behind the wall in an area, we first detect and track the Angles-of-Arrival (AoAs) of the signal reflections from these people, using a combination of 2D spectrum analysis, Joint Probabilistic Data Association Filter (JPDAF), and track management techniques.

- Given the AoAs corresponding to the different walking people in the area as a

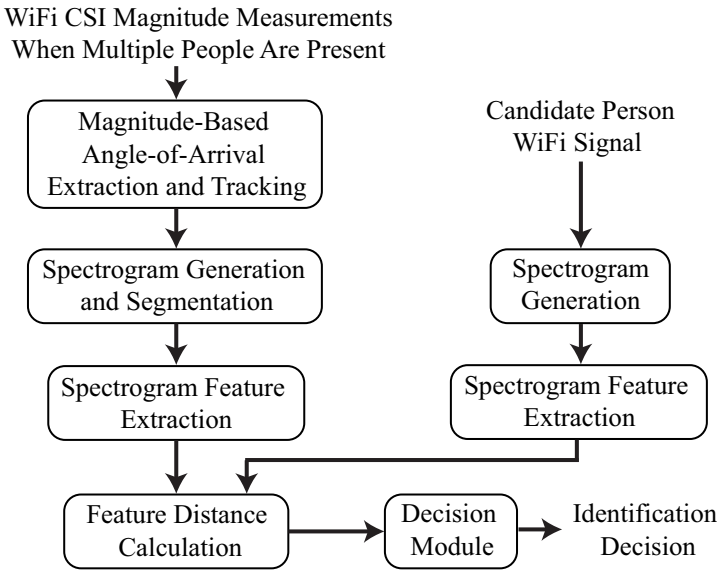


Figure 3.1: Flowchart showing the modules of our proposed system for multiple people identification.

function of time, we extract the gait signal of each walking person by projecting the total received signal to the AoA of this person. Next, given the separated WiFi signal of one person (referred to as the query), we calculate the spectrogram of this signal to capture the time-frequency characteristics of this person's gait.

- We then detect the segments from each spectrogram that are informative for identification. For each such segment, we extract several features to capture the person's gait signature.

- In order to determine whether a detected spectrogram segment (i.e., a query segment) belongs to the same person as the candidate person, we calculate the feature distances between a query segment and a candidate spectrogram, and feed them to a neural network which determines whether these two spectrograms correspond to the same person. As there can be multiple informative segments from the query spectrogram of a person's entire track during a measurement period, we use an algorithm based on maximum likelihood estimation to fuse the segment-based decisions.

The combined decision then indicates whether the query person of a walking track is the same as a candidate person.

3.1 Proposed Methodology

In this section, we describe our proposed methodology for multiple people identification. We first describe the multi-person identification scenario and the corresponding signal model. We then show how to estimate and track the Angles-of-Arrival (AoAs) of the reflections from the walking people, based on only the received signal magnitude measurements, and subsequently separate the signals relevant to each walking person by projecting the overall signal to the respective AoAs. Given the separated signal that contains the walk of one person, a frequency-time spectrogram is generated and the informative segments of it are extracted for identification.

3.1.1 Signal Model for Joint Processing Based on Only Magnitude Measurements

Consider the scenario shown in Fig. 3.2, where a WiFi transmitter (Tx) transmits WiFi signals that are reflected off of N people walking in an area with S static objects, and then received by a receiver array (Rx). The received signal is then a combination of the direct signal from the Tx to the Rx, the reflected signals off of the S static objects, and the reflected signals off of the N walking people. More specifically, the time-domain received signal as a function of the distance ℓ along the array, where ℓ is measured with

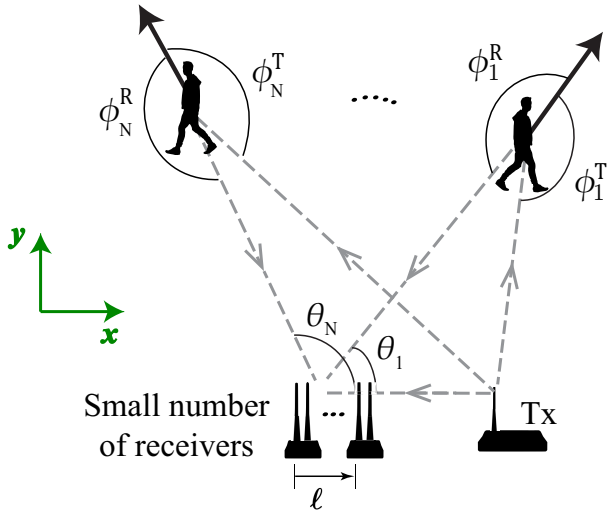


Figure 3.2: Multiple people are walking simultaneously in an area, where a WiFi transmitter and a small number of receivers (placed behind a wall, without a direct view of the people) are used to collect WiFi measurements. The black arrows in the figure indicate the respective walking directions of the people.

respect to the first antenna, can be written as follows [77]:

$$c(t,\ell) = \overbrace{\alpha_{\mathrm{D}} e^{-j\frac{2\pi}{\lambda}\ell\cos\theta_{\mathrm{D}}}}^{\text{Direct signal from Tx to Rx}} + \sum_{s=1}^{S} \overbrace{\tilde{\alpha}_s e^{-j\frac{2\pi}{\lambda}\ell\cos\tilde{\theta}_s}}^{\text{Reflected signal off of the s-th static scatterer}} + \sum_{n=1}^{N} \underbrace{\left(\sum_m \alpha_{n,m} e^{-j\frac{2\pi}{\lambda}v_{n,m}\psi_n^T t}\right) e^{-j\frac{2\pi}{\lambda}\ell\cos\theta_n}}_{\text{Reflected signals off of the n-th person}} + \eta(t,\ell), \tag{3.1}$$

where α_{D}, $\tilde{\alpha}_s$ and $\alpha_{n,m}$ are, respectively, the complex values of the direct path from Tx to Rx, the s-th static path (path reflected off of a static object), and the reflected path off of the m-th body part of the n-th person at $\ell = 0$, while θ_{D}, $\tilde{\theta}_s$, θ_n are their AoAs (measured with respect to the x-axis). S is the total number of static paths , $v_{n,m}$ is the speed of the m-th body part of the n-th person, $\psi_n^T = \cos\phi_n^T + \cos\phi_n^R$ is a parameter related to the motion direction of the n-th person (see Fig. 3.2 for definitions of ϕ_n^T and ϕ_n^R), $\eta(t,\ell)$ is additive noise, and λ is the wavelength of the WiFi signal.

Traditional AoA estimation relies on phase measurements across the receiver array.

In our case, however, we only have the received magnitude measurements. We next show how the AoA information can be preserved in the magnitude measurements as well. The squared magnitude of the received WiFi signal can be calculated from Eq. 3.1, as follows:

$$
\begin{aligned}
s(t) &= |c(t,\ell)|^2 = c(t,\ell)c^*(t,\ell) \\
&= P + \sum_{n=1}^{N}\sum_{m} 2|\alpha_{\mathrm{D}}\alpha_{n,m}|\cos\left(\frac{2\pi}{\lambda}\left(\psi_n^A\ell + v_{n,m}\psi_n^T t\right) + \angle\alpha_{\mathrm{D}} - \angle\alpha_{n,m}\right) \\
&+ \sum_{s=1}^{S} 2|\alpha_{\mathrm{D}}\tilde{\alpha}_s|\cos\left(\frac{2\pi}{\lambda}\tilde{\psi}_s^A\ell + \angle\alpha_{\mathrm{D}} - \angle\tilde{\alpha}_s\right) + \sum_{s}\sum_{s'\neq s}\tilde{\alpha}_s\tilde{\alpha}_{s'}^* e^{j\frac{2\pi}{\lambda}(\cos\tilde{\theta}_s - \cos\tilde{\theta}_{s'})\ell} \\
&+ \sum_{(n,m)}\sum_{s} 2|\alpha_{n,m}\tilde{\alpha}_s|\cos\left(\frac{2\pi}{\lambda}\left(v_{n,m}\psi_n^T t + (\cos\theta_n - \cos\tilde{\theta}_s)\ell\right) + \angle\alpha_{n,m} - \angle\tilde{\alpha}_s\right) \\
&+ \sum_{(n,m)}\sum_{\substack{(n',m') \\ (n',m')\neq(n,m)}} \alpha_{n,m}\alpha_{n',m'}^* e^{j\frac{2\pi}{\lambda}\left((v_{n,m}\psi_n^T - v_{n',m'}\psi_{n'}^T)t + (\cos\theta_n - \cos\theta_{n'})\ell\right)} + \eta'(t,\ell) \qquad (3.2)
\end{aligned}
$$

In Eq. 3.2, $P = |\alpha_{\mathrm{D}}|^2 + \sum_n\sum_m |\alpha_{n,m}|^2 + \sum_s |\tilde{\alpha}_s|^2$ denotes the total power of the signal, $\psi_n^A = \cos\theta_{\mathrm{D}} - \cos\theta_n$ is the difference between the cosine of the AoA of the direct path and the cosine of the AoA from the n-th person, $\tilde{\psi}_s^A = \cos\theta_{\mathrm{D}} - \cos\tilde{\theta}_s$ is the difference between the cosine of the AoA of the direct path from Tx to the Rx and the cosine of the AoA from the s-th static object, and $\eta'(t,\ell)$ is additive noise. Since the Tx and the Rx are close to each other, the direct path from Tx to Rx will be much stronger than the other reflected paths (i.e., $|\alpha_{\mathrm{D}}| \gg |\alpha_{n,m}|$ and $|\alpha_{\mathrm{D}}| \gg |\tilde{\alpha}_s|$). Hence, the first three terms in Eq. 3.2 dominate the rest of the terms, which can then be neglected. Furthermore, by subtracting the temporal mean of the received signal at each of the receiver antennas (which can be easily implemented in practice), all the time-independent terms of Eq. 3.2 are zeroed out. More specifically, after subtracting the temporal mean, the squared

magnitude of the received signal can be approximated as:

$$\begin{aligned}\bar{s}(t,\ell) &= |c(t,\ell)|^2 - \mathbb{E}_t\{|c(t,\ell)|^2\} \\ &\approx \sum_{n=1}^{N}\sum_{m} 2|\alpha_{\mathrm{D}}\alpha_{n,m}|\cos\left(\frac{2\pi\left(\psi_n^A\ell + v_{n,m}\psi_n^T t\right)}{\lambda} + \mu\right) + \eta'(t,\ell),\end{aligned} \tag{3.3}$$

where $\mu = \angle\alpha_{\mathrm{D}} - \angle\alpha_{n,m}$, and $\mathbb{E}_t\{.\}$ denotes the temporal mean of the argument. The first term in Eq. 3.3 is a superposition of N terms, each of which corresponds to one of the moving persons, and contains a combination of sinusoids whose frequencies depend on the AoA of that person to the receiver (ψ_n^A) and the speeds of that person's body parts ($v_{n,m}\psi_n^T$). This shows that $\bar{s}(t,\ell)$ carries vital information about the AoAs of the reflected signals off of the walking people.

By taking the 2D Fourier Transform of $\bar{s}(t,\ell)$ with respect to time t and distance ℓ, we then have

$$\begin{aligned}C(f_T, f_D) &= \left|\iint \bar{s}(t,\ell)e^{-j\frac{2\pi}{\lambda}(f_T t + f_D \ell)}dt\,d\ell\right| \\ &= \sum_{n=1}^{N}\sum_{m}|\alpha_{\mathrm{D}}\alpha_{n,m}|\delta\left(f_T - v_{n,m}\psi_n^T, f_D - \psi_n^A\right) \\ &\quad + \sum_{n=1}^{N}\sum_{m}|\alpha_{\mathrm{D}}\alpha_{n,m}|\delta\left(f_T + v_{n,m}\psi_n^T, f_D + \psi_n^A\right) + N_\circ,\end{aligned} \tag{3.4}$$

where f_T is the temporal frequency, f_D is the spatial frequency, $\delta(*,*)$ is the 2D Dirac Delta function, and $N_\circ$ is the noise floor.

It can be seen from Eq. 3.4 that each reflector in the operation area results in 2 deltas in the 2D spectrum $C(f_T, f_D)$. By placing the Tx such that the AoA of the direct path from Tx to Rx is zero (i.e., $\cos\theta_{\mathrm{D}} = 1$), the values of $\psi_n^A = 1 - \cos\theta_n$ are restricted to the interval $[0, 2]$, which is disjoint from the range of values of $-\psi_n^A$ [77]. Therefore, by constraining the search space for ψ_n^A to $[0, 2]$ (e.g., using only positive values for f_D), we

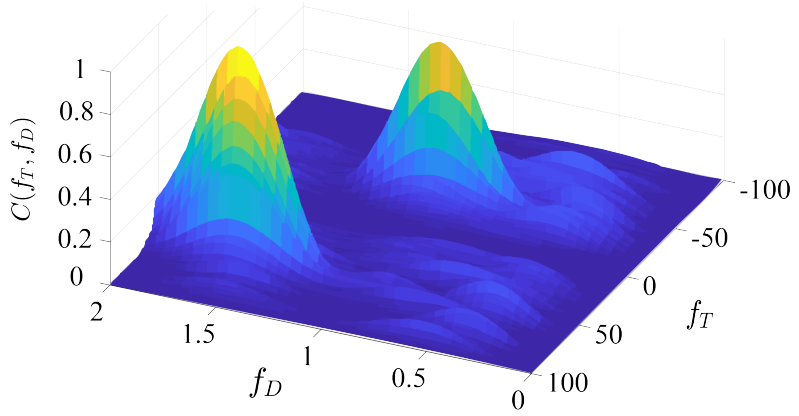


Figure 3.3: A sample 2D spectrum in the (f_T, f_D) domain for the case where two people are walking simultaneously in an area, based on a 0.4-second WiFi measurement. It can be seen that there are two peaks appearing in the 2D spectrum, corresponding to the strong torso reflections of the two walking people.

obtain only one delta function for each reflector in the area at $(f_T, f_D) = (v_{n,m}\psi_n^T, \psi_n^A)$ in the 2D spectrum. For the delta functions of the body parts of the same person, the one with the largest coefficient corresponds to the strongest reflecting part of the body, which is the torso. We refer to this strongest delta of a person as a peak and denote its location in the 2D spectrum by $(\bar{\psi}_n^T, \psi_n^A)$. Fig. 3.3 shows an example of a 2D spectrum $C(f_T, f_D)$ generated from WiFi measurements collected using an 8-element antenna array in a real experiment, where two people are walking simultaneously in an area. The ground-truth AoAs of the two people are 90° and 125° (i.e., $\psi_1^A = 1$ and $\psi_2^A = 1.57$), respectively. It can be seen that there are two peaks at $(f_T, f_D) = (-25, 1)$ and $(f_T, f_D) = (54, 1.6)$ in the 2D spectrum, which correspond to the strong torso reflections of the two walking people.

Our new 2D model can additionally detect whether a person is walking towards or away from the link. As can be seen from Fig. 3.2, the reflected signal from a person walking away from the link has a negative ψ^T, while that from a person walking towards the link has a positive ψ^T. The peak location in the 2D spectrum then indicates the walking direction, as can be seen in Fig. 3.3, where the first person (whose $\bar{\psi}^T = -25$)

is walking away from the link while the second person (whose $\bar{\psi}^T = 54$) is walking towards it. Hence, even if two people have the same AoA at the receiver array but are walking in different directions, their reflections are separable in the 2D spectrum $C(f_T, f_D)$. On the other hand, in the traditional 1D signal model, the peak corresponding to a walking person appears twice, at $\pm v_m \psi^T$ at each time instance, which makes the motion direction of the walking person ambiguous, unless the information of all the subcarriers are exploited.

Based on the properties of the Fourier Transform, we summarize the following properties about the shape of $C(f_T, f_D)$ in the f_D dimension:

- The width of a peak in the f_D dimension is inversely proportional to the length of the Rx antenna array. Thus, to increase the resolution (or separability) of the measurements in the AoA domain, one can increase the length of the array.

- Based on the Nyquist sampling theorem, the maximum allowable antenna separation in the antenna array is determined by the maximum possible ψ^A. More specifically, the maximum antenna separation is given by: $\Delta\ell = \lambda/(2\psi^A_{\max})$. Therefore, if the AoAs of the reflected signals from the walking people span the range of [0, 180°], the maximum allowable antenna separation is $\lambda/4$.

Let $C_t(f_T, f_D)$ be the 2D spectrum generated from a time window of the WiFi measurements from t to $t+T_{\text{win}}$. We then measure the temporal sequence of peaks appearing in consecutive snapshots of $C_t(f_T, f_D)|_{t=0:\Delta t:T_{\max}}$ (where $T_{\max}$ is the total measurement duration and Δt is the time step) to keep track of the AoAs of different people, as we show next.

3.1.2 AoA Tracking

Given the 2D spectrum $C_t(f_T, f_D)$ generated from a time window of the received WiFi signal from t to $t + T_{\text{win}}$, we extract a set of J_t peaks located at $\psi_j = (\bar{\psi}_j^T, \psi_j^A)$, for $j \in \{1, ..., J_t\}$. We denote the set of peak locations at this time instance by $\mathbf{\Psi_t} = \{\psi_j : j = 1, \ldots, J_t\}$. In order to separate and extract the part of the received signal that is relevant to each of the walking people in the area, the system needs to perform two main tasks as described below.

1. AoA Track Management: A track $\mathcal{T}$ is defined as a temporal sequence of peaks in the 2D spectrum resulting from the same moving person. As we do not assume that the system has any prior knowledge about the number of people in the area or when each person starts walking, it is necessary to automatically initialize a track when a new person starts walking in the area, or terminate a track when the corresponding person stops walking or leaves the area. We refer to a track for a person currently walking in the area as an *active track*, a newly initialized but not yet confirmed track (due to the possibility of false alarm) as a *potential track*, and a terminated track for which the person does not walk in the area anymore as a *finished track*.

2. AoA Data Association: At each time instance, given a set of active and potential tracks and the set of new peaks in the 2D spectrum, the system needs to automatically associate each peak with the track of the corresponding walking person.

In order to perform these tasks, we propose an AoA tracking algorithm, as described in Alg. 1. The algorithm takes as input the measurement sequence, in the form of the set of peaks in the 2D spectrum over time. We then utilize the Joint Probabilistic Data Association Filter (JPDAF) to associate the peaks with the current active or potential tracks (lines 3 and 4 in Alg. 1), with filter parameters pertaining to the association process estimated from the training measurements. We refer the reader to [78] for more details on

Algorithm 1 Angle-of-Arrival Tracking Algorithm

Input: Set of peaks in the 2D spectrum over time: $\mathbf{\Psi_t}|_{t=0:\Delta t:T_{\max}}$
Output: Set of active tracks and finished tracks: $\mathcal{A}$, $\mathcal{F}$

1: Initialize the sets of active tracks $\mathcal{A}$, potential tracks $\mathcal{P}$, and finished tracks $\mathcal{F}$ to empty sets: $\mathcal{A} = \emptyset$, $\mathcal{P} = \emptyset$, $\mathcal{F} = \emptyset$.
2: **for** $t = 0 : \Delta t : T_{\max}$ **do**
3: Find the association between the current set of peaks $\mathbf{\Psi_t}$ in the 2D spectrum $C_t(f_T, f_D)$ and the current active and potential tracks using the JPDAF.
4: For each track $\mathcal{T} \in \mathcal{A} \cup \mathcal{P}$ that is associated with a peak in $\mathbf{\Psi_t}$, update its $\bar{\psi}^T$ and ψ^A using the associated peak.
5: For each peak $\psi_j \in \mathbf{\Psi_t}$ that is not associated with any active or potential tracks in $\mathcal{A} \cup \mathcal{P}$, initialize a new potential track $\mathcal{T}_{\text{new}}$ with ψ_j and add it to the current set of potential tracks: $\mathcal{P} = \mathcal{P} \cup \{\mathcal{T}_{\text{new}}\}$.
6: **for** each $\mathcal{T} \in \mathcal{P}$ **do**
7: **if** the potential track $\mathcal{T}$ meets confirmation criteria **then**
8: $\mathcal{T}$ is confirmed as an active track and is then moved to the set of active tracks: $\mathcal{P} = \mathcal{P} \backslash \mathcal{T}$, $\mathcal{A} = \mathcal{A} \cup \{\mathcal{T}\}$.
9: **else**
10: $\mathcal{T}$ is a false alarm and is then deleted from the set of potential tracks: $\mathcal{P} = \mathcal{P} \backslash \mathcal{T}$.
11: **end if**
12: **end for**
13: **for** each $\mathcal{T} \in \mathcal{A}$ **do**
14: **if** the active track $\mathcal{T}$ meets the termination criteria **then**
15: remove $\mathcal{T}$ from $\mathcal{A}$ and add it to the set of finished tracks: $\mathcal{A} = \mathcal{A} \backslash \mathcal{T}$, $\mathcal{F} = \mathcal{F} \cup \{\mathcal{T}\}$.
16: **end if**
17: **end for**
18: **end for**

JPDAF. During the association, a measured AoA closer (in terms of Euclidean distance) to the current AoA of a track is more likely to belong to that track, as compared to an AoA that is farther away.

Once the JPDAF associates the peaks/measurements to the current set of tracks, our proposed algorithm enforces the following set of rules for track management:

- Initialization criterion (line 5 in Alg. 1): A potential track is initialized if a peak in the 2D spectrum is not associated with any active or potential tracks at the current time instance.

- Confirmation criterion (line 7 in Alg. 1): A potential track is confirmed active if it is associated with some peaks in the 2D spectrum for more than a percentage p_{T_C} of the time during the first T_C seconds after initialization. Otherwise, it is considered as a false alarm and discarded.

- Termination criterion (line 14 in Alg. 1): An active track is finished if it is not associated with any peaks for more than a percentage p_{T_F} of the time during the past T_F seconds.

Then, the total number of people in the area, N, will be estimated as the total number of active and finished tracks at the end of the experiment. We can then obtain the AoA time series of the n-th person: $\hat{\theta}_n(t) = \cos^{-1}(1 - \psi_n^A(t))$, where $\psi_n^A(t)$ is the time series of ψ^A in the n-th track $\mathcal{T}_n$. Fig. 3.4 shows a sample output of the AoA tracking algorithm for a 40-second experiment of two people walking. The track of person A is initialized at $t = 2.2$ s. The AoA time series of person A is shown by the blue solid curve in Fig. 3.4. Similarly, the track of person B is initialized at $t = 5.1$ s, with the corresponding AoA time series shown by the red dashed curve in the figure. For each track, the AoA sequence $\hat{\theta}_n(t), n = 1, \ldots, N$, is utilized to generate the spectrogram that only contains the gait information of the corresponding person, isolated from the signals of the other walking people, as we show next.

3.1.3 Spectrogram Generation and Segmentation

Given the time series of the AoA of the reflected path off of a walking person, we project the received signal at the Rx array, $\bar{s}(t, \ell)$, to the AoA of this person. The projected signal contains only the reflections from this person, and thus makes it possible to analyze this person's gait and perform identification. We then perform time-frequency analysis on the projected signal. More specifically, given $\hat{\psi}_n^A(t)$, the spectrogram for the

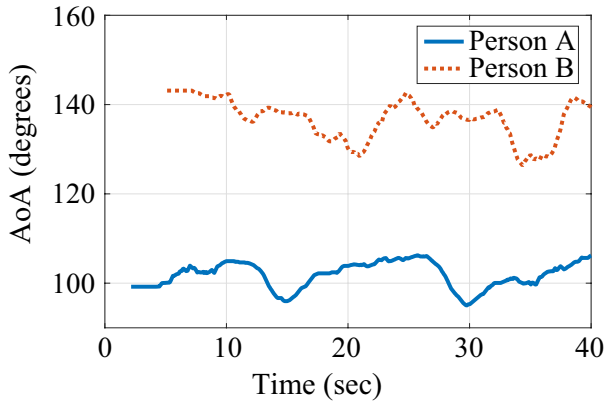


Figure 3.4: Sample output of the AoA tracking algorithm. Measurements of person A are first detected at $t = 2.2$ s (blue solid curve). Measurements of person B are first detected at $t = 5.1$ s (red dashed curve). Both tracks remain active till the end of the experiment.

n-th walking person is generated by

$$S_n(t,f) = \left| \int_\tau \underbrace{\left(\int_\ell \bar{s}(\tau,\ell) e^{-j\frac{2\pi}{\lambda}\hat{\psi}_n^A(\tau)\ell} d\ell \right)}_{\substack{\text{projecting the signal} \\ \text{to the n-th person's AoA}}} \chi(\tau;t) e^{-j2\pi f\tau} d\tau \right|^2, \tag{3.5}$$

where $\hat{\psi}_n^A(t) = 1 - \cos\hat{\theta}_n(t)$ and $\chi(\tau;t)$ is a time window function starting at time t. In other words, the measured signal $\bar{s}(\tau,\ell)$ across all the Rx antennas as a function of time is first projected to the n-th person's AoA, using the operation inside the inner brackets in Eq. 3.5. This results in a one-dimensional signal (as a function of time) which contains only the reflected signal off of the n-th person. This one-dimensional signal is then time-windowed using a window $\chi(\tau;t)$ and the frequency content of the time-windowed signal (which contains that person's gait information) is obtained using FFT. Different window functions $\chi(t)$ result in different favorable characteristics in the spectrogram. For instance, as shown in the literature, a rectangular window (in which case $S_n(t,f)$ is the Short-Time Fourier Transform of the projected signal) produces a spectrogram that has clearer information on the speeds of the limbs, whereas the Hermite window functions

produce spectrograms with better time-frequency resolution [30]. As such, we utilize a multi-window spectrogram by additively combining the spectrograms generated using different window functions, as follows:

$$S_n(t,f) = \sum_{k=0}^{K} \left| \int_{\tau} \left(\int_{\ell} \bar{s}(\tau,\ell) e^{-j\frac{2\pi}{\lambda}\hat{\psi}_n^A(\tau)\ell} d\ell \right) \chi_k(\tau;t) e^{-j2\pi f\tau} d\tau \right|^2,$$

where $\chi_0(\tau;t)$ is the rectangular window function starting at time t, and $\chi_k(\tau;t)$ is the k-th Hermite function [68].

As an illustrative example for our spectrogram generation method, Fig. 3.5 (left) shows the spectrogram of the overall measured WiFi signal squared magnitude in a real experiment when two people are walking simultaneously in an area. It can be seen that the gait signals of the two people severely interfere with each other, making it extremely challenging to identify them. On the other hand, by utilizing our AoA tracking and spectrogram generation approach, we obtain two separate spectrograms (Fig. 3.5 (right)), where each spectrogram carries the gait information of only one individual.

When a person is walking towards/away from the WiFi link, the value of ψ^T is approximately constant when the Tx and Rx are close to each other. Hence, the frequency components of the spectrogram $S_n(t,f)$ directly correspond to the speeds of different body parts of the n-th person (see Eq. 3.4) and we do not have to consider the walking route to compensate for ψ^T.[1] Furthermore, in Chapter 2, we have shown that the *most informative segments* of a spectrogram are those that correspond to the person walking away from the WiFi link as the reflection from the back of the body results in a cleaner signal. We then utilize the spectrogram segmentation algorithm in Chapter 2 to extract these informative segments (where ψ^T is constant and the person is walking away from the link) from the spectrogram of the entire walking duration of a person, with the same

[1] See Sec. 2.1 and 2.5 for more detailed discussions on ψ^T.

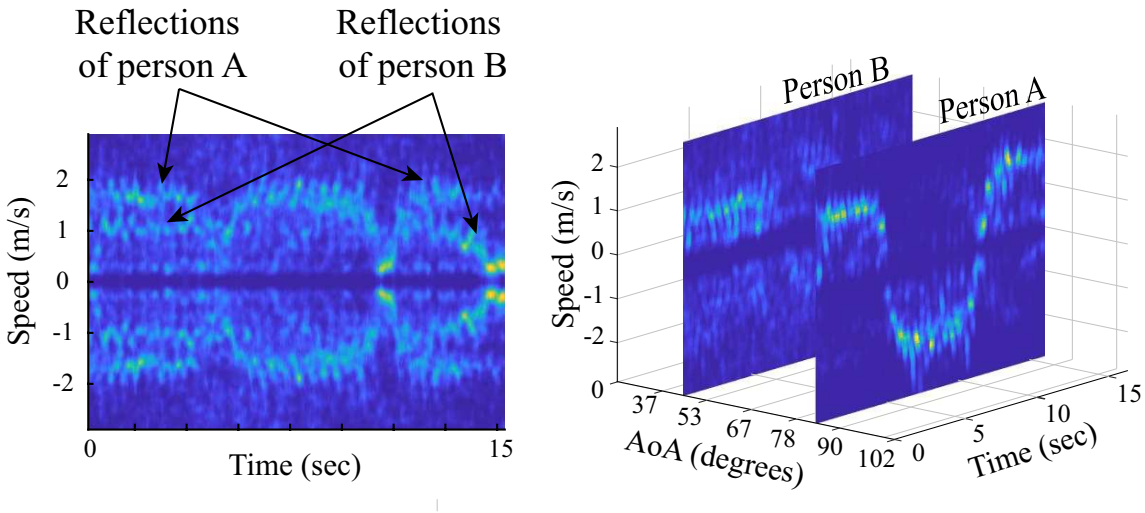


Figure 3.5: (Left) Sample spectrogram of the received WiFi signal when two people are walking simultaneously in an area, showing that their gait signatures are mixed up in the overall signal. (Right) The output of our 2D signal processing pipeline, showing two separate spectrograms each carrying the gait information of only one person.

algorithm parameters as in Chapter 2 . The subjects can walk on any general path. The segmentation algorithm will then extract only the informative segments of each person's walk. Note that there can be more than one such informative segments from a person's track, depending on the duration and the path.

In case the reflected signals from two (or more) people are not separable for a long time, e.g., due to them always walking very close to each other, a corresponding extracted spectrogram segment could contain the mixed gait information of both people. Such segments, which are not useful for identification, can be automatically detected and removed, since their energy spread over frequency is very large due to the combined gait patterns. More specifically, we consider the energy spread of a segment obtained during operation to be abnormally large if it is 20% more than the maximum energy spread observed in the training spectrograms generated from single-person walking scenarios. The energy spread of a spectrogram is measured by the 80th percentile of the difference between the 60th and 40th percentiles of the frequency distribution across time. Addi-

tionally, when two people are walking very closely, it is also possible that a spectrogram segment of one person gets dominated by the gait information of the other person (e.g., due to a possibly stronger reflection from the person who is close-by). Such a segment can be automatically detected and discarded, based on its very high similarity to the segments of another person/track. The final remaining spectrogram segments are then considered as the *valid segments* to be used for identification.

3.2 Feature Extraction and Person Identification

Given a spectrogram segment of a person's walk derived from the separated WiFi signal, our proposed system identifies this person by comparing this query segment with a candidate spectrogram. More specifically, we extract several features from both the query segment and the candidate spectrogram, and then compute a set of distances between them. Given these feature distances, we train a neural network to determine whether a pair of query segment and candidate spectrogram belong to the same person. In this way, the neural network is able to identify people who have not been seen during training. Furthermore, if multiple valid segments are obtained from a person's track, we can fuse the neural network's decisions on these segments via a maximum likelihood approach.

3.2.1 Spectrogram Features and Distances

We extract 10 key features from a spectrogram to capture a person's gait. More specifically, we look at the frequency and time dimensions of the spectrogram, which carry different types of gait information that are useful for identification. The frequency dimension carries information about the speeds of different body parts, as shown in Eq. 3.3. We use the following features to capture various aspects of the frequency infor-

mation:

• **Frequency Distribution:** This feature is obtained by averaging the spectrogram over time. This feature captures the average distribution of the frequency components, i.e., the speed distribution of the body parts, during the person's walk.

• **Frequency Distribution in Four Gait Phases:** These are the time averages of the spectrogram for each of the 4 phases of the gait cycle [1]. This feature provides more detailed information on the speed profile of the body parts of the walking person in each gait phase.

• **Average Leg and Torso Speeds:** We calculate the averages of the torso speed curve and the leg speed curve, respectively, over time. These speed curves can be extracted from the spectrogram using the method in [1].

In addition, we use several time-domain features to capture the temporal information on a person's gait (e.g., the gait cycle) as follows:

• **Average Autocorrelation:** Given a spectrogram, we compute the autocorrelation across time (with a maximum lag of 2 sec) for each frequency bin. We then compute a weighted sum of the autocorrelation curves of the frequency bins, based on the energy distribution over the frequencies. This autocorrelation-based feature captures the periodic pattern of the speed profile of the person's body during walking.

• **Autocorrelation of Percentile Curves:** Given a spectrogram, we extract the 50th and 70th percentiles of the frequency distribution over time. We then compute the autocorrelation of these two percentile curves, respectively. These features also capture the periodicity of the gait and are more robust to noise.

In order to quantitatively compare the similarity between the query spectrogram segment and the candidate spectrogram, for each of the 10 spectrogram features, we compute the distance between the corresponding features of the query spectrogram segment and the candidate spectrogram, which results in a vector of 10 feature distances.

More specifically, for the frequency distributions, we use the Kullback-Leibler Divergence (KLD) as the distance metric, which is commonly used to compare distributions. For the autocorrelation-based features, we use the cosine similarity, which compares both the values and patterns of the autocorrelations. For the average speed features, we use the Euclidean distance.

3.2.2 Identification

Given a pair of a query spectrogram segment and a candidate spectrogram, we compute the 10 feature distances as described previously. We then utilize a simple fully-connected neural network, which has 1 hidden layer with 30 units, to combine these distances into a binary identification decision. More specifically, the neural network takes as input a 10-dimension vector consisting of the feature distances and outputs a binary classification decision indicating whether or not the person of the query segment from the multiple people scenario is the same as the person of the candidate spectrogram. The neural network is trained on spectrogram pairs of subjects and locations disjoint from those of the test set (more details on the training set in Sec. 3.3.3). During training, these 10 distances between a pair of spectrograms of the training subjects are fed into the neural network, along with a binary label indicating whether these two spectrograms belong to the same person. A positive label of 1 indicates that the pair of spectrograms belong to the same person and a negative label of 0 indicates otherwise. Utilizing the softmax operation, the neural network outputs a scalar soft decision $\xi \in [0, 1]$ that resembles the probability of declaring a same-person spectrogram pair given the input, i.e., $\xi = p(1|\text{input})$. The neural network is trained using the cross-entropy loss and does not suffer from overfitting due to its simple structure. After training, we estimate the distribution of the values of the output ξ on all the positive-label training data, $p(\xi|1)$, as

well as its distribution on all the negative-label training data, $p(\xi|0)$. We then estimate a Likelihood Ratio function: $LLR(\xi) = p(\xi|1)/p(\xi|0)$.

During the testing phase, we first compute the distances between the previously described 10 features of the query segment (a valid spectrogram segment of a person among the group of walking people) and the candidate spectrogram. This 10-dimension distance vector is then fed into the trained neural network, for identifying whether or not the query spectrogram segment and the candidate spectrogram belong to the same person. The neural network outputs a soft decision ξ, for which we calculate $LLR(\xi)$ using the LLR function estimated from the training set. If $LLR(\xi) > 1$, a positive binary decision is declared that the person of the query spectrogram segment is the same as the person of the candidate spectrogram. Otherwise, the system declares that the query segment and the candidate spectrogram belong to two different people.

3.2.3 Multi-Segment Decision Fusion

As previously discussed in Sec. 3.1.3, there can be multiple valid spectrogram segments for the same track/person in an experiment. Let N_{VS} be the number of valid segments for the n-th person from a group of walking people. Let ξ_i, $i \in \{1, ..., N_{\text{VS}}\}$, be the soft decision outputs when the i-th valid segment is tested against the candidate spectrogram. To optimally combine these soft decisions, in terms of maximum likelihood, we adopt the following decision rule: $p(\xi_1, \ldots, \xi_{N_{\text{VS}}}|1) \underset{0}{\overset{1}{\gtrless}} p(\xi_1, \ldots, \xi_{N_{\text{VS}}}|0)$. Furthermore, assuming independent decisions by the neural network on the different N_{VS} inputs, the decision rule becomes: $\prod_{i=1}^{N_{\text{VS}}} LLR(\xi_i) \underset{0}{\overset{1}{\gtrless}} 1$, where the query person is declared to be the same as the one of the candidate spectrogram if the product of the LLR values corresponding to the multiple segments is greater than 1.

3.3 Experimental Setup and Data Collection

In this section, we discuss our experiments used to validate our proposed system.

3.3.1 Experiment Setup

For the WiFi data collection, we use laptops equipped with the Intel 5300 WiFi Card to act as WiFi transmitter and receiver. For the transmitter, we use a tripod-mounted antenna connected to one port of the Intel card in one laptop, which transmits 1,000 WiFi packets per second on WiFi channel 36 (with a center frequency of 5.18 GHz). For the receiver array, we use 4 laptops. More specifically, we use 8 antennas in a linear placement (with antenna separation of $\lambda/4$, except for one area, for which we use $\lambda/2$ separation), mounted to a tripod 1 m away from the Tx and connected to the ports of the Intel WiFi cards of 4 laptops, where each laptop provides 2 antenna ports. Since we rely only on the CSI magnitude measurements of the received WiFi signals, no calibration or synchronization between the multiple NICs is needed. This also facilitates increasing the antenna array length if needed. The CSI measurements of the receiver laptops are logged using Csitool [79] and processed offline using our proposed algorithm in Sec. 3.1. For spectrogram generation, we use a time window of length $T_{\text{win}} = 0.4\,\text{s}$ and a time shift of 4 ms.

The Intel 5300 WiFi cards report the CSI measurements on 30 different subcarriers. We utilize the extra measurements in the subcarrier domain to denoise the signals. More specifically, we utilize Principal Component Analysis (PCA) to find the top 5 PCA components of the subcarrier measurements, and sum up their spectrograms in order to generate the denoised spectrogram. As opposed to traditional PCA denoising [1], where the measurement on each subcarrier is a 1D temporal signal, the measurement on each subcarrier in our case is a 2D spatiotemporal signal $c(t, \ell)$. Hence, we utilize the

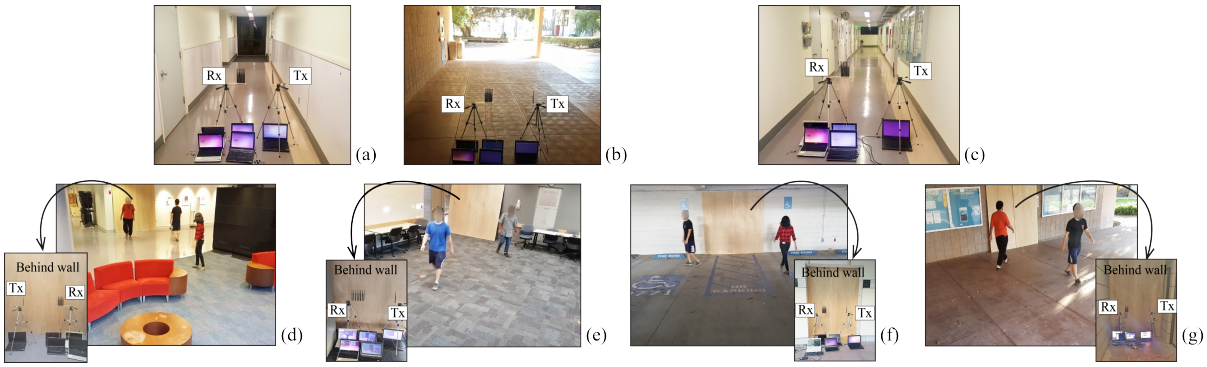


Figure 3.6: (a)–(b) WiFi training areas: We collect the WiFi measurements of the training subjects in a line-of-sight setting in these two areas. (c) Test candidate area: We collect the measurements of the test subjects in this hallway area, in a line-of-sight setting, to be used for the candidate spectrograms. (d)–(g) Test areas: We conduct the test experiments in these 4 behind-wall areas (lounge, conference room, parking lot, and outdoor area), which represent a variety of real-world scenarios. The WiFi transmitter and receiver are placed behind the wall in each area, as shown by the black arrow.

multidimensional PCA to find the top 5 PCA components [80].

For the AoA track management criteria (discussed in Sec. 3.1.2), we set the parameters as follows: $p_{T_C} = 50\%$, $T_C = 2$ sec, $p_{T_F} = 90\%$, and $T_F = 10$ sec.

3.3.2 Training Experiments

Training Subjects: We have recruited a total of 13 training subjects (10 male subjects and 3 female subjects) to collect data for training the neural network described in Sec. 3.2.2. The training subjects have an average walking speed of 1.22 m/s (standard deviation: 0.19 m/s) and an average gait cycle of 1 s (standard deviation: 0.1 s).

Training Areas and Data Collection: The training data are collected in the two areas shown in Fig. 3.6 (a)–(b), in a line-of-sight (non-through-wall) setting. For each training subject, we collect 9 WiFi measurements of their walk in each of the two training areas. For each measurement, the training subject walks away from the link for about 10 m. We then transform each WiFi measurement into a corresponding spectrogram,

which results in a total of 234 spectrograms for training.

Training Dataset: As our neural network of Sec. 3.2.2 is designed to predict whether a query spectrogram segment and a candidate spectrogram belong to the same person, we construct the training set in the form of spectrogram pairs. More specifically, based on the 234 training spectrograms, we generate a total of $27,261$ spectrogram pairs. For each pair, we calculate the distances between the corresponding features of the two spectrograms and use this distance vector as one training sample. A training sample is given a positive label if the pair of spectrograms belong to the same training subject and a negative label otherwise. Since we have different numbers of positive and negative training samples, we utilize oversampling to obtain a balanced training set [73].

3.3.3 Test Experiments

Given the WiFi measurements when a person was walking in an area (referred to as the candidate person), and the WiFi measurements when multiple people are walking behind the wall in another area (test area), our system can separate the gait information of each person in the test area and determine if he/she (referred to as the query person) is the same as the candidate person. Both query and candidate people were not seen during training. In this part, we describe the details of our evaluation setup, including the test subjects, the data collection of the candidate spectrograms, and the test experiments.

Test Subjects: Our test set has 6 subjects (5 male subjects and 1 female subject), none of whom is part of the training set. Their heights range from 162 cm to 186 cm. They have an average speed of 1.31 m/s (standard deviation: 0.27 m/s) and an average gait cycle of 1.01 s (standard deviation: 0.15 s).

Candidate Spectrograms: We collect 10 WiFi measurements for each test subject in the area of Fig. 3.6 (c) in a line-of-sight (non-through-wall) setting to serve as candidate

spectrograms. For each measurement, one test subject walks away from the link for about 10 m. The WiFi measurements are transformed into spectrograms, which then serve as the candidate spectrograms for identification. As a WiFi measurement in this setting only involves one person walking away from the link, each candidate spectrogram translates into one valid segment that continuously spans the entire walking duration. Thus, instead of using the term "candidate spectrogram segment", we use the shortened form "candidate spectrogram" in this chapter.

Test Experiments: We conduct the test experiments in 4 different test areas, as shown in Fig. 3.6 (d)–(g), which are disjoint from both the training areas of Fig. 3.6 (a)–(b) and the area where the candidate spectrogram was generated. These test areas represent several real-world scenarios with a variety of area size, geometry, and clutterness. More specifically, the area of Fig. 3.6 (d) is a lounge area with couches and coffee tables. The area of Fig. 3.6 (e) is a conference room with several tables and chairs. The area of Fig. 3.6 (f) is located inside a parking structure. The area of Fig. 3.6 (g) is a roofed outdoor area near a building. We conduct a total of 92 test experiments in these test areas, where the WiFi Tx and Rx are placed behind walls, with no direct view of the walking subjects. The walls consist of wooden panels that attenuate the wireless signals by 4 to 5 dB, which is similar to or larger than those of common non-concrete and non-metal building materials [81]. It is worth noting that wood is used for the walls of 90% of the residential and small commercial buildings in the U.S. [71].

First, consider the test experiments with 2 people (total of 80 experiments). In each such experiment, 2 of the 6 test subjects are randomly selected to walk simultaneously in the area. Each subject then walks casually on general paths. During the experiment, the AoAs of the walking subjects are tracked at the receiver, and their respective individual spectrograms are separated and extracted from the aggregate WiFi signal as described in Sec. 3.1. The informative and valid segments of the spectrograms of the tracks are

automatically detected as discussed in Sec. 3.1.3, which are then used as the query segments. In the test experiments with 3 people, 3 of the 6 test subjects are randomly selected in a total of 12 experiments. The AoAs of the 3 walking people are then tracked to extract individual spectrograms, whose valid segments then serve as the query segments.

Test Dataset for the Case of 2 People: This test set consists of the test subjects' candidate spectrograms and the query spectrogram segments from the test experiments when two people were present, in the form of query-candidate pairs. More specifically, each test pair consists of a candidate spectrogram and a query spectrogram segment, and is assigned a positive label if they belong to the same person and a negative one otherwise. Our test set has in total 15,792 such test pairs.

Since there can be more than one query spectrogram segment obtained from the same person's track in an experiment, we also look at the test cases where the system identifies whether the person of a track is the same as the person of a candidate spectrogram, by fusing the decisions of all the valid segments from this track as discussed in Sec. 3.2.3. In this setting, we have a total of 4,892 pairs of a query track and a candidate spectrogram. We refer to the first setting where each query is one single segment as the *segment-based* setting and the second setting where the query consists of all the valid segments from a track of a person as the *track-based* setting.

Test Dataset for the Case of 3 People: For the test experiments with 3 people walking simultaneously, the corresponding test set has 2,416 query-candidate test pairs in the segment-based setting and 1,512 test pairs in the track-based setting.

3.4 System Evaluation

In this section, we present extensive experimental evaluations of our proposed system in 4 different through-wall areas (see Fig. 3.6), and for both cases of 2 and 3 people in

Area	Two people		Three people	
	Segment based	Track based	Segment based	Track based
Area of Fig. 3.6 (d)	79%	81%	79%	73%
Area of Fig. 3.6 (e)	83%	82%	79%	83%
Area of Fig. 3.6 (f)	84%	86%	81%	87%
Area of Fig. 3.6 (g)	77%	78%	79%	83%
Average	**81%**	**82%**	**80%**	**83%**

Table 3.1: The segment-based and track-based identification accuracies of our proposed system on the test set, in 4 different through-wall areas. The last row shows the average performance over all the areas.

the area. Furthermore, the subjects and areas in the test experiments have never been seen during the training phase.

3.4.1 Evaluation Criteria

We evaluate our proposed system by using pairs of a query spectrogram segment and a candidate spectrogram. Given such a test pair, the system identifies whether they belong to the same person or not. The resulting binary classification accuracy is used as the evaluation metric. As we have different numbers of test pairs with positive (same-person) labels and negative (different-people) labels, we report the balanced classification accuracy, i.e., the average of the respective accuracies over the same-person and different-people pairs.

3.4.2 Evaluation with Two Walking People

We evaluate our proposed system on our extensive test set, as described in Sec. 3.3.3, which contains only areas and subjects not seen during training. The results are summarized in Table 3.1. For the track-based metric, our proposed system achieves identification accuracies of 81%, 82%, 86%, and 78% for the 4 areas, respectively, with an overall aver-

age accuracy of 82%. This indicates that given the spectrogram segments of a person's track in an experiment, our system is able to correctly identify whether this person is the same as the test subject of a candidate spectrogram for 82% of the time. As for the segment-based setting, our system achieves identification accuracies of 79%, 83%, 84%, and 77% for the 4 areas, respectively, with an overall average of 81%.

Finally, we show the Receiver Operating Characteristic (ROC) curve for the track-based identification in Fig. 3.7. It can be seen that the ROC curve is significantly above the 45-degree line. In particular, the Area Under Curve (AUC) is 0.89 in the track-based setting (and 0.88 in the segment-based setting), indicating a good performance. The AUC is equal to the probability that a randomly-drawn positive sample has a higher score (e.g., the $LLR(\xi)$ in Sec. 3.2.2) than a randomly-drawn negative sample.[2] It should be noted our threshold is automatically given by the neural network as described in Sec. 3.2.2 and 3.2.3. These results further confirm the performance of our system.

3.4.3 Angular Separation

We analyze the system's performance with respect to the angular separation between two walking people (difference between their AoAs) in the test areas. Fig. 3.8 shows the average identification accuracy as a function of the angle. It can be seen that even when the angle between the two people is as small as 20°,[3] our system still performs robustly with an accuracy similar to those when the angular separation is larger.

[2]See [82] for more details on ROC and AUC.

[3]Note that the minimum angular separation resolvable by the system depends on the length of the Rx array. For instance, given our current hardware setup, the system will not be able to fully resolve two angles that are 10° (or less) apart. See Sec. 3.5 for a detailed discussion on AoA resolution.

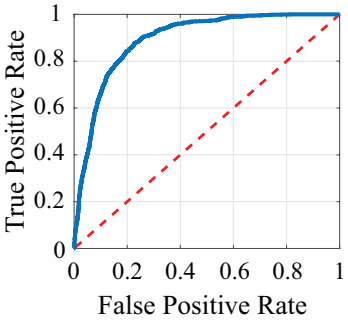


Figure 3.7: Receiver Operating Characteristic (ROC) curve for identification in the track-based setting.

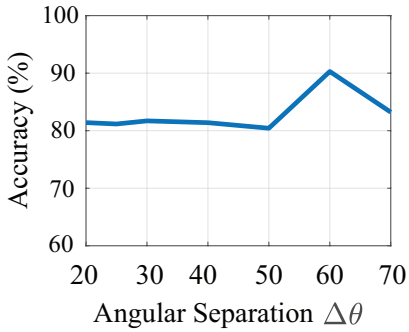


Figure 3.8: Track-based identification accuracy as a function of the angular separation between the subjects' tracks in the two-people experiments.

3.4.4 Evaluation with Three Walking People

For the three-people test experiments, our proposed system achieves overall segment-based and track-based identification accuracies of 80% and 83%, respectively, averaged over the 4 through-wall test areas. It can be seen that even with 3 people simultaneously walking in the area, our system still achieves a good identification accuracy.

Overall, these extensive test results show that our proposed system can successfully perform identification when multiple people are simultaneously walking in the area, even though the people and areas have never been seen during training, and in a through-wall setting. They thus demonstrate the efficacy and applicability of our system in various real-world scenarios.

3.5 Discussions

In this section, we provide discussions on various aspects of our proposed framework, as well as possible future extensions.

Gait as a Unique Signature: While some studies suggest that gait identification errors are inevitable for large groups (e.g., 100 people) [83], other studies have shown

that RF-based gait identification can be effective for small to medium group sizes (e.g., up to 50 people) [1, 2]. It should be noted, however, that these papers have the same subjects in training and test, which makes the learning problem considerably simpler. They nonetheless show that gait-based identification has a good potential for applications that involve a small to medium pool of subjects (e.g., in a residential or an office setting).

Resolution of Peaks in the AoA Domain: The maximum number of people that the system can simultaneously detect is determined by the resolution (or width) of a peak in the AoA domain, which depends on length of the RX array. In our experiments, we used an antenna array of total length 2λ, resulting in a first-null peak resolution of 0.23, i.e., two peaks are completely separated in the AoA domain if the difference of their ψ^A values is greater than 0.23. This resolution can further be improved by extending the array, which does not add any synchronization/calibration overhead since we use only the magnitude of the received WiFi signal. Different processing techniques (e.g., MUSIC) can also improve the resolution of the peaks, at the expense of much higher computational costs.

Increasing the Number of People in the Area: While we have tested our system for up to 3 people walking simultaneously in an area, the proposed approach can be used for a larger number of people. The maximum number of people that the system can identify is determined by the resolution of a peak in the AoA domain, which depends on the length of the RX antenna array, as discussed above, as well as how crowded the area gets, how close to each other the people walk, and their walking directions. For instance, in our current setup, the AoA resolution is 10 degrees. This means that if two people consistently walk with AoA separation of less than 10 degrees in the same direction, our current setup cannot separate them (if they walk in different directions, then it is still separable). As such, as long as the area is large enough such that each person's AoA is not consistently less than 10 degrees from someone else who is going in

the same direction, then the current system is able to identify all the people. As the area gets more crowded for its size, then one can increase the length of the array to allow for a better AoA resolution, which would allow our system to identify people even if they are consistently walking close to each other in the same direction. One can also utilize more resources, such as positioning more transceivers in different locations in the space in order to capture different views of the people, and/or utilizing different frequency subcarriers to capture the time-of-flight data.

Tracks of the Walking People: In Sec. 3.1.3, we have utilized the fact that when people move away/towards the link, ψ^T is constant, excluding the need to directly estimate it via tracking people's routes. In the rare case when no such segment is detected in the entire spectrogram (e.g., if the person's whole walking path is parallel to the link), the constantly varying value of ψ^T can be estimated by existing RF-based tracking systems, and then incorporated in the proposed approach to identify a person walking on any general track.

Impact of Misalignment of the Rx Array: When the Rx array is rotated by θ_e, the AoA of a person changes from θ to $\theta + \theta_e$, where θ is the AoA before the rotation. The separability of the reflected signals from two people depends on the difference between their respective ψ^A values, where $\psi^A = \cos(\theta_e) - \cos(\theta + \theta_e)$. As such, given a small θ_e, the separation between two people in the 2D spectrum can slightly increase or decrease due to the nonlinear cosine operation. For instance, consider two people whose original AoAs are 120° and 150°. When the array is rotated by 5°, the separation between their ψ^A values decreases from 0.37 to 0.33. Note that given our current hardware setup, the two people's signals are fully separable in the 2D spectrum as long as their ψ_A values are at least 0.23 apart. Our system is also robust to small inter-antenna spacing errors, since they only slightly affect the width of the main beam and the side lobe levels, but not the location/AoA of the main beam [84].

Chapter 4

Teaching RF to Sense Without RF Training Measurements

In this chapter, we describe our proposed general framework for translating the already-available online video content for motion-based activities to the RF domain, in order to generate instant RF data for training human-motion-based RF sensing systems. More specifically, given the video footage of a person performing an action (e.g., walking, running, physical exercises), our framework allows one to simulate the RF signal that would have been measured if this person performed the action in the vicinity of RF transceivers. As such, our proposed framework makes it possible to tap into the virtually unlimited publicly-available online videos to generate instant wireless data and construct a massive simulated RF dataset for training RF sensing systems for various applications.

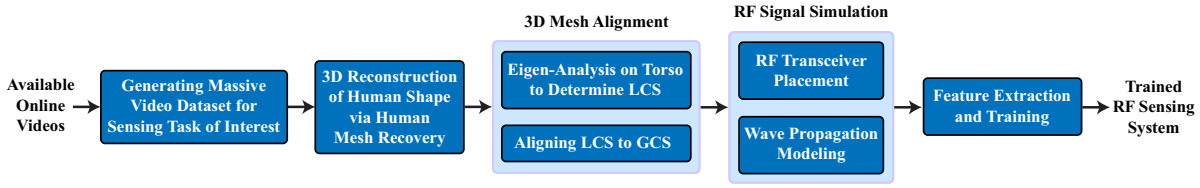


Figure 4.1: Various steps involved in our proposed framework for training RF sensing systems solely based on video data. LCS stands for Local Coordinate System, while GCS stands for Global Coordinate System.

4.1 Video-based Training for RF Sensing Systems

Fig. 4.1 summarizes the key steps of our proposed framework. We first collect a massive video dataset pertaining to the RF sensing task of interest, using available online videos. For each video of a person doing an activity, our pipeline first extracts a 3D mesh of this person as a function of time. This extracted 3D human model is then transformed into a Global Coordinate System (GCS) that is independent of the camera view using eigen analysis. Then, given the desired RF sensing configuration (e.g., transceiver positions, frequency), our framework simulates the RF signals that would have been measured if the extracted human mesh was in the given RF sensing setup, via efficient wave propagation modeling. Through time-frequency analysis, we then extract key features from this simulated RF dataset and train an RF sensing system, which will then be deployed in a real wireless environment during operation. We next discuss each of these components in detail.

4.1.1 3D Reconstruction of Human Shape

The first step towards translating video content to the wireless domain is to build an accurate 3D shape of the person in a video frame. In order to do so, we utilize the recent advances in computer vision [62, 85, 86, 87]. For instance, [62] proposes a Human Mesh Recovery (HMR) algorithm that uses a convolutional encoder network and a regression

module to infer information about the pose and the shape of a person from a single 2D image. Then, they utilize the Skinned Multi-Person Linear (SMPL) model [88] to translate the inferred body pose and shape information into a *human mesh*. An SMPL human mesh is a set of triangulated points in 3D space describing the body surface. More specifically, the extracted 3D human mesh is characterized by a dense set of mesh points (e.g., 6890 points) that describe, in 3D, the outer surface of the human body in the video. Fig. 4.2 (a) shows examples of our reconstructed 3D human mesh from a snapshot of a person doing jumping jacks and stiff-leg deadlifts, using the algorithm of [62]. In the first step of our approach, we then extract the 3D human shape from the video, using the state-of-the-art mesh reconstruction algorithms in computer vision.

4.1.2 3D Mesh Alignment via Eigen-analysis

The goal of our proposed framework is to simulate the RF signals that would have been measured by one or more RF transceivers if the person in the video was in their vicinity, given any RF *sensing setup*. By sensing setup, we mean the relative location and orientation of the person with respect to the RF transceivers, as well as the relative locations among the transceivers. We thus need to correctly place the transceivers with respect to the extracted human mesh in the simulation environment, in order to generate a given sensing setup, which requires translating the mesh to a global coordinate system.

In the reconstructed 3D human mesh from the output of a vision algorithm (e.g., [62]), the coordinates of the 3D points are calculated based on the camera view, which is estimated from the given video frame. In other words, the reconstructed mesh may reside in different coordinate systems across different videos, when the videos are shot from different views. Therefore, it is essential to transform the 3D mesh into a Global Coordinate System (GCS) that is invariant to the camera view. Such a transformation

Figure 4.2: (a) Sample reconstructed 3D human mesh from a snapshot of (left) a person doing jumping jacks and (right) a person doing stiff-leg deadlifts. (b) The Local Coordinate System (LCS), defined with respect to the human body.

allows us to arbitrarily choose the orientation and the position of the human mesh in the target GCS, as well as place the RF transceivers in the GCS as needed for the sensing setup. In order to enable this transformation, however, we need to know some information about the orientation of the body parts. We next show how to achieve this through an eigen-analysis on the mesh points of the torso.

Let the set of 3D points of the extracted human mesh at time t be $\mathcal{M}'(t) = \{\mathbf{p'}_m(t), m \in \{1, \ldots, M\}\}$, where $\mathbf{p'}_m(t) \in \mathbb{R}^3$ is the 3D location of the m-th point at time t and M is the total number of points in the mesh. These points are given in a coordinate system where the $\mathbf{x}$-$\mathbf{y}$ plane is parallel to the camera plane. We define a Local Coordinate System (LCS) with respect to the human body, where the $\mathbf{x}$-axis points to the front of the person, the $\mathbf{y}$-axis points to the person's left, and the $\mathbf{z}$-axis points upward, as shown in Fig. 4.2 (b). The axes of the LCS are denoted by $\mathbf{x}'$, $\mathbf{y}'$, and $\mathbf{z}'$, respectively, and are determined based on the set of mesh points $\mathcal{M}'$ as follows. Let $\mathbf{H} = [\mathbf{p'}_{s_1}, \mathbf{p'}_{s_2}, \ldots, \mathbf{p'}_{s_{M_T}}]$ be a $3 \times M_T$ matrix of the 3D locations of all the mesh points belonging to the torso in the original coordinate system, where $s_1, \ldots, s_{M_T}$ are the indices of the torso points among the M total mesh points and M_T is the total number of torso points.[1] Since the

[1]The indices of the mesh points belonging to any specific body part are fixed and known for all meshes generated by a human mesh recovery algorithm.

anterior-posterior (i.e., front-back) is the smallest dimension of the human torso, $\mathbf{x}'$ is the eigenvector of $\mathbf{H}\mathbf{H}^\intercal$ that corresponds to the smallest eigenvalue, where $^\intercal$ is the transpose operator. Similarly, $\mathbf{z}'$ is the eigenvector of $\mathbf{H}\mathbf{H}^\intercal$ that corresponds to the largest eigenvalue, since the inferior-superior (i.e., bottom-top) is the largest dimension of the human torso.

Once we have determined the axes $\mathbf{x}'$, $\mathbf{y}'$, and $\mathbf{z}'$ of the LCS corresponding to the human mesh of a video frame, the human mesh can be rotated such that the person faces any desired direction in the GCS. For instance, consider the case where the person is required to face the positive $\mathbf{x}$-axis in the GCS in the simulation, we multiply all the original mesh points by the rotation matrix $\mathbf{R} = [\mathbf{x}', \mathbf{y}', \mathbf{z}']$, i.e., $\mathbf{p}_m(t) = \mathbf{R}^\intercal \mathbf{p}'_m(t)$. It is straightforward to show that after this operation, the LCS of the rotated mesh points $\mathbf{p}_m(t)$ aligns with the $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ axes of the GCS. The mesh points can also be easily translated in the new GCS to any arbitrary location. For example, if the $\mathbf{x}$-$\mathbf{y}$ plane in the GCS is assumed to be the floor and the person is in a standing position, we can translate the mesh points such that the feet points lie on the $\mathbf{x}$-$\mathbf{y}$ plane.

In general, depending on the configuration of the RF sensing setup, such as the transceiver positions relative to the person, the human mesh can be put into any desired orientation and location in the GCS. Concrete examples will be provided in our case study of Sec. 4.2.

4.1.3 RF Signal Simulation

Let $\mathcal{M}(t) = \{\mathbf{p}_m(t), m \in \{1, \ldots, M\}\}$, where $\mathbf{p}_m \in \mathbb{R}^3$ is the 3D location of the m-th point of the human mesh in the GCS, where the values of $\mathbf{p}_m(t)$ were calculated such that the human model has a specific location and orientation in the GCS, as described previously. Let $\mathbf{p}_T$ and $\mathbf{p}_R$ denote the locations of the RF transmitter (Tx) and receiver

(Rx) in the GCS, respectively, in the sensing setup of interest. We then simulate the RF signal that would have been measured from reflections off of the extracted human mesh in this setup.

In general, the received electric field at the Rx, $E(\mathbf{p}_R)$, due to a transmission from the Tx, is the solution to the volume integral equation [89]:

$$E(\mathbf{p}_R) = \mathbf{g}(\mathbf{p}_R, \mathbf{p}_T) + \iiint_{\mathbf{D}} \mathbf{g}(\mathbf{p}_R, \mathbf{p}) O(\mathbf{p}) E(\mathbf{p}) d\mathbf{p}, \tag{4.1}$$

where $\mathbf{D}$ is the workspace where the transceivers and the person are located, $\mathbf{g}(\mathbf{p}_1, \mathbf{p}_2) = \frac{\exp(j\frac{2\pi}{\lambda}\|\mathbf{p}_1 - \mathbf{p_2}\|)}{4\pi\|\mathbf{p}_1 - \mathbf{p_2}\|}$ is the Green's function from point $\mathbf{p}_2$ to point $\mathbf{p}_1$, where $\|.\|$ denotes the Euclidean distance and λ is the wavelength of the wireless signal. $O(\mathbf{p})$ is a parameter that captures the electric/magnetic properties of what resides at position $\mathbf{p}$ in the area. In the electromagnetics literature, several methods have been proposed for finding the solution to Eq. 4.1, such as the Method of Moments (MoM) [90] and the Finite Element Method (FEM) [91]. However, they are very computationally intensive. Thus, efficient linearizing approximations to Eq. 4.1 have been proposed, such as the Born approximation [64]. Based on our extensive studies (reported in detail in Sec. 4.2.4), Born approximation well approximates the details of the received signal with the accuracy needed for training based on the simulated RF data, while being computationally very efficient. In Born approximation, we have the following for the received signal:

$$E(\mathbf{p}_R) = \mathbf{g}(\mathbf{p}_R, \mathbf{p}_T) + \iiint_{\mathbf{D}} \mathbf{g}(\mathbf{p}_R, \mathbf{p}) O(\mathbf{p}) \mathbf{g}(\mathbf{p}, \mathbf{p}_T) d\mathbf{p}, \tag{4.2}$$

where the total electric field at point $\mathbf{p}$, $E(\mathbf{p})$, inside the integral of Eq. 4.1 is approximated by the Green's function from the Tx to point $\mathbf{p}$. This means that the Born approximation treats each point in space as an independent scatterer, and does not take

into account the higher-order scattering effects.

In summary, given a video of a person engaged in some activity, we first extract a dense set of 3D mesh points describing the outer human surface. We then transform the reconstructed 3D human mesh into a GCS, which can be put into any arbitrary desired location and orientation in the GCS. Then, based on the given sensing setup, we determine the Tx and Rx positions in the GCS, and simulate the RF signal that would have been measured using the approximated Born wave model of Eq. 4.2. More advanced wave models (e.g., Eq. 4.1) can certainly be used as part of the proposed pipeline if more details are needed for a particular application, at the cost of higher computational complexity. Furthermore, given a different RF sensing setup, we can easily re-run the simulation to obtain the corresponding RF signal, by changing the transceiver locations or the orientation and location of the already-aligned human mesh in the GCS, according to the new setting. As we shall see in Sec. 4.2, the proposed pipeline can generate realistic RF data for the purpose of training an RF sensing system.

4.1.4 Feature Extraction and Training

Once the RF signals are simulated based on the desired scenario, they can be used to train a machine learning algorithm for a given RF sensing application. For traditional machine learning algorithms, such as support vector machine and neural network, one can extract several features from the simulated RF signals for training the system. Since our proposed framework enables the generation of massive RF training data, it is also possible to utilize our framework to generate sufficient data for training deep learning algorithms.

Overall, our proposed scalable and general framework enables training RF sensing systems without the need for collecting any real RF training data, and by translating

the vast available video data to the RF domain. Given the generated RF training data, one can then apply any machine learning algorithm to train the RF sensing system. In the next section, we then showcase the possibilities created by the proposed framework, in the context of RF sensing for gym activity recognition.

4.2 Case Study: Gym Activity Classification

In this section, we demonstrate the efficacy of our proposed framework with a real-world application of WiFi-based activity recognition. More specifically, we consider gym activity classification, the objective of which is to identify the performed activity from a set of several different physical exercises, such as push-up and sit-up, using only WiFi CSI magnitude measurements of a small number of links. RF-based gym activity classification is a challenging problem due to the complex movements involving different parts of the body, which has only been explored sparsely in the literature [28, 42, 43]. However, all such existing work on gym activity classification require a significant effort in collecting massive wireless training measurements for the set of activities that they want to classify, using the same RF-sensing setup that will be used during the operation phase.

In contrast to these existing papers, we show how to train a WiFi-based gym activity classifier without the need for collecting any wireless training measurements. As discussed earlier, our proposed framework enables the translation of the video content of a human activity into the wireless domain, which allows us to utilize the available online videos to create an instant RF training dataset. In this section, we then discuss our WiFi sensing setup for gym activity classification, as well as how we implement the different steps of our proposed framework for this real-world application.

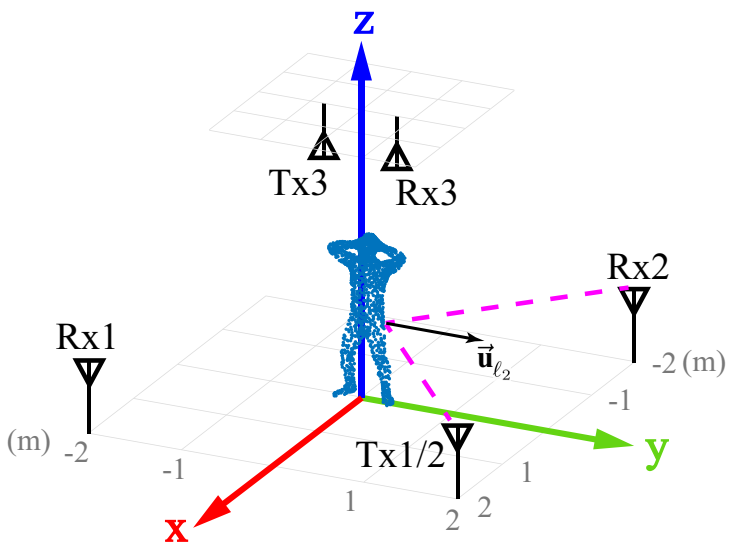


Figure 4.3: The sensing setup used for our case study of gym activity classification. The 3 WiFi links capture the velocity components of different body parts along the 3 dimensions.

4.2.1 Sensing Setup

In this case study, the WiFi sensing system is tasked with capturing the characteristics of different gym activities to perform classification. As such, we consider a sensing setup that is capable of measuring the motion profile of each activity, which can then serve as the signature for classification.

Consider the setup of Fig. 4.3, with the coordinate system as marked, where the person is located at the origin, facing the positive **x** direction, and the positive **z**-axis is pointing upward. Our WiFi sensing system consists of 3 links, each with a pair of transmitter (Tx) and receiver (Rx). Link 1 is placed in front of the person and parallel to the **y**-axis, Link 2 is placed to the left of the person and parallel to the **x**-axis, while Link 3 is placed above the person and parallel to the **y**-axis, as shown in Fig. 4.3. Links 1 and 2 share the same Tx, which we denote by Tx1/2. As the person performs the activity, the WiFi signal emitted by the Tx bounces off of the person's body, as well as the static objects in the environment, and is received by the Rx, for each link. The

baseband WiFi signal received by the Rx of the i-th link can be approximated as follows:

$$c_i(t) \approx c_i^{\circ} + \sum_s c_i^{\mathrm{s}} + \sum_n \underbrace{\alpha_n e^{j\frac{4\pi}{\lambda}\langle \vec{\mathbf{v}}_n(t)\cdot \vec{\mathbf{u}}_{\ell_i}\rangle t}}_{\substack{\text{reflected signal off} \\ \text{the n-th body part}}}, \tag{4.3}$$

where c_i° is the direct signal from the Tx to the Rx of the i-th link, c_i^{s} is the signal of s-th static path (path reflected off of a static object in the environment), α_n is the amplitude of the reflected path off of the n-th body part, $\vec{\mathbf{v}}_n(t)$ is the 3D velocity vector of the n-th body part at time t, $\vec{\mathbf{u}}_{\ell_i}$ is a unit vector bisecting the angle between the two lines connecting the n-th body part to the Tx and Rx of the i-th link, $\langle * \cdot * \rangle$ is the inner product of the vector arguments, and λ is the wavelength of the RF signal. We are interested in gym activity recognition using only WiFi CSI magnitude measurements. In practice, the direct path from the Tx to the Rx is stronger than all reflected paths (i.e. $|c_i^{\circ}| \gg |c_i^{s}|$, $|c_i^{\circ}| \gg \alpha_n$) due to their longer lengths and the reflection losses. Hence, the squared signal magnitude can be written as [67],

$$\begin{aligned}
|c_i(t)|^2 = P &+ \sum_n 2|c_i^{\circ}|\alpha_n \cos\left(\frac{4\pi}{\lambda}\langle \vec{\mathbf{v}}_n(t)\cdot \vec{\mathbf{u}}_{\ell_i}\rangle t - \angle c_i^{\circ}\right) \\
&+ \sum_n \sum_s 2\alpha_n |c_i^{\mathrm{s}}| \cos\left(\frac{4\pi}{\lambda}\langle \vec{\mathbf{v}}_n(t)\cdot \vec{\mathbf{u}}_{\ell_i}\rangle t - \angle c_i^{\mathrm{s}}\right) \\
&+ \sum_n \sum_{n'>n} 2\alpha_n \alpha_{n'} \cos\left(\frac{4\pi}{\lambda}\langle (\vec{\mathbf{v}}_n(t) - \vec{\mathbf{v}}_{n'}(t))\cdot \vec{\mathbf{u}}_{\ell_i}\rangle t\right) \\
\approx P &+ \sum_n 2|c_i^{\circ}|\alpha_n \cos\left(\frac{4\pi}{\lambda}\langle \vec{\mathbf{v}}_n(t)\cdot \vec{\mathbf{u}}_{\ell_i}\rangle t - \angle c_i^{\circ}\right),
\end{aligned} \tag{4.4}$$

where $P = |c_i^{\circ}|^2 + \sum_s \sum_{s'} |c_i^{\mathrm{s}}||c_i^{\mathrm{s'}}| e^{j\left(\angle c_i^{\mathrm{s}} - \angle c_i^{\mathrm{s'}}\right)} + \sum_s 2|c_i^{\mathrm{s}}||c_i^{\circ}| \cos\left(\angle c_i^{\mathrm{s}} - \angle c_i^{\circ}\right) + \sum_n \alpha_n^2$ is the DC component of $|c_i(t)|^2$, $\angle$ denotes the phase of the signal, and the approximation in the last line of Eq. 4.4 is due to the fact that the direct path is stronger than the reflected ones. It can be seen from Eq. 4.4 that the static multipath in the environment affects

the DC component of the received signal, which does not carry any motion information (i.e., information on $\vec{\mathbf{v}}_n(t)$), and can be easily subtracted in practice.

Given the sensing setup of Fig. 4.3, we can see that $\vec{\mathbf{u}}_{\ell_1}$, $\vec{\mathbf{u}}_{\ell_2}$, and $\vec{\mathbf{u}}_{\ell_3}$ can be approximated by $[1,0,0]^\intercal$, $[0,1,0]^\intercal$, and $[0,0,1]^\intercal$, respectively. For instance, Fig. 4.3 shows $\vec{\mathbf{u}}_{\ell_2}$ in our setup, which, as can be seen, is approximately a unit vector along $\mathbf{y}$. $\langle \vec{\mathbf{v}}_n \cdot \vec{\mathbf{u}}_{\ell_i} \rangle$, $\forall i = 1,2,3$, are then the 3 components of the velocity vector $\vec{\mathbf{v}}_n$ in the 3D space along the three directions of the Cartesian coordinate axes. Then, based on Eq. 4.4, the frequency content of the received signals at the 3 links will directly capture the velocity components of different body parts along the $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ directions, respectively. As such, the received signals at the 3 links contain key information on the person's 3D motion profile, which will be very useful for the classification task.[2]

We next demonstrate how to translate relevant online video data to WiFi training data using our proposed framework, for the gym activity classification task. We start by describing our implementation of the various steps shown in Fig. 4.1, tailored for the gym activity recognition and for the sensing setup of Sec. 4.2.1.

4.2.2 Training with Zero RF Training Data

In this case study, the gym activity classification includes 10 different physical exercises, such as jumping jack and push-up, as shown in Fig. 4.4. They are representative of a variety of typical workouts that involve the movements of different body parts.

We download the training videos for these activities from YouTube.[3] For each video, we manually identify and extract the overall chunk in which the person is actually performing the corresponding activity (to remove chunks where, for instance, the person is

[2]Other configurations of the links can be translated to the setup of Fig. 4.3 by projecting their measured motion information along the 3 axes in Fig. 4.3, as we shall discuss in Sec. 4.3.

[3]Here are the links to a few sample training videos for interested readers: `https://youtu.be/96zJo3nlmHI` (broad jump), `https://youtu.be/UpyDdQjBTa0` (forward lunge), `https://youtu.be/33UV3Jl8wEk` (jumping jack), and `https://youtu.be/FvJS_MSN4Lo` (lateral lunge).

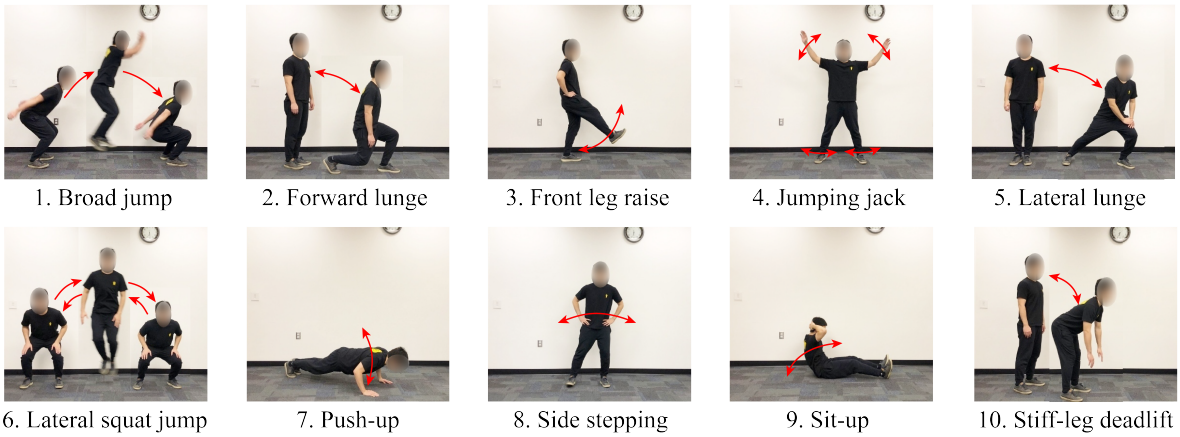


Figure 4.4: The set of 10 activities considered in our gym activity classification case study. In our study, the activities are performed with only body weight, i.e., without any equipment (e.g., resistance band, dumbbell). These images show the movements involved in each activity, as indicated by the red arrows. See the color pdf for optimal viewing.

just talking). Although this is done manually in the current study, it is possible to use recent computer vision techniques to temporally localize such activity periods automatically [92, 93]. In order to best facilitate the 3D mesh extraction, we do not use videos in which there is blockage of the exercising person by the surrounding objects. This is not a restrictive requirement since most online videos of gym activities provide unobstructed views of the person.[4] Our training video dataset contains a total of 61 videos of gym activities, each of which containing an average of 8.39 sec of relevant activity content.

Activity Repetition Segmentation

In our WiFi-based gym activity classification system, a repetition is taken as the atomic unit to describe the gym activities. More specifically, a repetition is defined as one complete movement cycle of a gym activity, such as one push-up or one jumping jack.

Given a video of a person performing a gym activity for multiple repetitions (e.g.,

[4]While it is ideal to use videos without any occlusion, the vision algorithms are still able to reconstruct the 3D human mesh if the occluded part is small, e.g., [62, 94].

doing 10 jumping jacks), we want to segment the time duration for each individual repetition. In order to do so, we first use Mask R-CNN [95] to extract the bounding box of the person for each frame. As the person performs an activity for multiple repetitions, the shape of the bounding box as a function of time captures this periodicity and can be used to calculate the time duration for each repetition. For instance, in a video of a person doing jumping jacks, the height and width of the bounding box varies periodically due to the arm and leg movements. We then segment the individual jumping jacks based on the autocorrelation function of the aspect ratio of the bounding box, which is changing with time.

3D Human Mesh Extraction and Alignment

For each video frame of the activity, we use the HMR algorithm of [62] to extract the 3D mesh of the person, which consists of a large number of 3D points describing the outer surface of the person in the image. We then use the alignment framework proposed in Sec. 4.1.2 to align the extracted 3D mesh of the person to the 3D Global Coordinate System (GCS) of the setup described in Sec. 4.2.1. While we can, in principle, align any given frame using the method of Sec. 4.1.2, some frames are easier to align as they would require the extraction of the least amount of information from the image, pertaining to the positions of different body parts. We next show our approach to find such a frame for each activity video, in the context of the stiff-leg deadlift exercise.

Consider the stiff-leg deadlift exercise, in which a person slowly lowers their upper body from the standing position to a bend-forward position, and then quickly rises back to the standing position, while keeping their legs straight (see Fig. 4.4). Given any random frame, the angle between the person's torso and legs needs to be estimated from the frame, in order to determine the correct corresponding orientation in the GCS, and consequently calculate the rotation matrix. While this knowledge can be estimated from

each frame, if we use the frame where the person is fully standing, we do not need to extract any additional knowledge in order to build the rotation matrix, which can then be applied to all the frames of the video. In order to find a frame where the person is in a fully standing position during the stiff-leg deadlift exercise, we utilize the Mask-RCNN algorithm [95] to estimate a bounding box around the person in each frame, which can then automatically identify the frame of the person fully standing as the one with the tallest bounding box. According to our sensing setup of Fig. 4.3, the LCS axes $\mathbf{x}', \mathbf{y}', \mathbf{z}'$ of the person in the frame with fully standing position should align with the major $\mathbf{x}, \mathbf{y}, \mathbf{z}$ of the GCS. We can then easily calculate the rotation matrix $\mathbf{R}$ from the LCS of this frame as $\mathbf{R} = [\mathbf{x}', \mathbf{y}', \mathbf{z}']$, and use this matrix to align the meshes of all the frames of this video into the correct orientation in the GCS. Since the feet are static in this activity, the mesh should be translated in the GCS such that the average of the mesh points of the feet is at the origin.

As an additional example, consider the sit-up exercise. The fully recumbent position is the easiest to use for alignment. More specifically, given the sensing setup of Sec. 4.2.1, when the person is in the full recumbent position, $\mathbf{x}'$, $\mathbf{y}'$, and $\mathbf{z}'$ of the LCS should align with $+\mathbf{z}$, $+\mathbf{y}$, and $-\mathbf{x}$ axes of the GCS, respectively. Hence, the rotation matrix $\mathbf{R}$ can be estimated from the LCS of the video frame in which the person is fully lying down as: $\mathbf{R} = [-\mathbf{z}', \mathbf{y}', \mathbf{x}']$.

As such, we then use the common-sense knowledge of each activity to provide a label for a key frame that our automated algorithm needs to look for, in order to efficiently align the meshes of all the video frames of that activity to the GCS.

WiFi Signal Simulation

After the mesh alignment, the 3D mesh of the person doing the activity is placed and oriented in the GCS of Fig. 4.3 in a simulation environment. We further place the WiFi

transceivers in the locations described in Sec. 4.2.1 in our simulation environment, where the Tx-Rx separation distances for Links 1 and 2 are 4 m each, and the Tx-Rx distance for Link 3 is 0.6 m. The antennas of Links 1 and 2 are placed at 0.75 m above the floor, and the antennas for Link 3 are placed at 2.75 m above the floor (which is a typical height of room ceilings). As the human mesh moves over time, we simulate the received WiFi signals for all the links as a function of time, by utilizing the Born electromagnetic approximation (Eq. 4.2). More specifically, given the locations of the M 3D mesh points in the GCS at any time instant (any video frame), we simulate the received WiFi signal as follows,

$$c_v(\mathbf{p}_R; t) = \left| \mathbf{g}(\mathbf{p}_R, \mathbf{p}_T) + \sum_{m=1}^{M} A_m G_m \, \mathbf{g}\Big(\mathbf{p}_R, \mathbf{p}_m(t)\Big) \, \mathbf{g}\Big(\mathbf{p}_m(t), \mathbf{p}_T\Big) \right|^2, \tag{4.5}$$

where $\mathbf{p}_R$ and $\mathbf{p}_T$ are the locations of the Rx and Tx, respectively, $\mathbf{p}_m(t)$ is the location of the m-th mesh point at time t, A_m is the reflection coefficient of the m-th mesh point, and G_m is a scaling parameter that captures the quasi-specular reflection nature of the human body and depends on the normal direction to the body at the m-th mesh point. Since the clothing has a negligible impact on the reflection coefficient, we model human body as a homogeneous reflector (constant A_m). Furthermore, since we are only interested in the motion-related part of the received signal (see Eq. 4.4), we do not need to calculate the exact value of the received signal and a scaled version will suffice to model human motion. As such, we use a uniform constant reflection coefficient of 1 over the whole body.

Remark 4.1 *The reflected signals off of the static objects in a real environment, e.g. walls and furniture, contribute to the DC term of Eq. 4.4, which carries no information about the human motion, and can be easily removed in the operation phase, as we shall see in Sec. 4.2.3. Hence, there is no need to consider the static multipath in the simulation*

environment and we only need to consider the mesh points of the moving human body in the received signal calculation in Eq. 4.5.

When using the HMR algorithm of [62], particularly for video frames where one of the person's arms is occluded from camera view, we have observed some artifacts in the estimation of the arm poses, which may result in abnormal arm movements that the person in the video did not perform. We also noticed that due to its small surface area, as compared to the other body parts, the arms have little contribution to the received WiFi signal [72]. For these reasons, we do not consider the mesh points of the arms in the simulation platform, and only model the interaction between the electromagnetic waves and the rest of the human body.

Remark 4.2 *At higher frequencies (e.g., 60 GHz) or for some other applications, the impact of arms on the received signal could be higher. In such a case, one can then include the mesh points of the arms in the simulation to capture the impact of the arms on the received signal. If one needs to include the arms in the modeling, other more recent vision algorithms that better estimate the arms can be used [94]. Alternatively, instead of YouTube videos, one can use existing online motion capture data which directly provide high-quality 3D human models.*

Feature Extraction and Classification

For each gym activity, the body parts of the person would produce a different velocity profile, in terms of the speed and the motion direction. Such a velocity profile, i.e., the information about the speed and motion direction of different body parts, can be used as a signature for each gym activity. For instance, when doing push-ups, the up-down motion of the person's body parts (e.g., head, torso) results in moderate speeds in the $\pm\mathbf{z}$ direction of the GCS. Meanwhile, push-up produces nearly no speeds in the $\mathbf{x}$ and

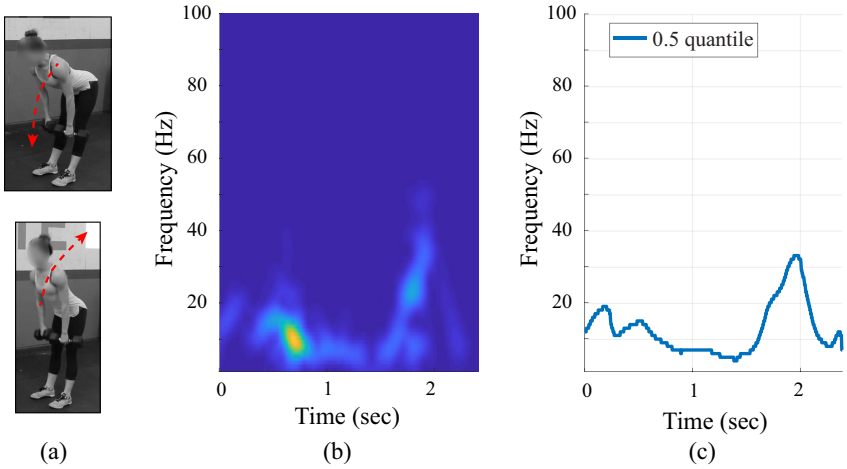


Figure 4.5: (a) Snapshots from a video of a person performing a repetition of the stiff-leg deadlift exercise, (b) the corresponding spectrogram of the simulated WiFi signal of Link 3 capturing the motion pattern of the person, and (c) the 0.5 quantile curve of the spectrogram in (b). See the color pdf to best view this figure.

$\mathbf{y}$ directions. As another example, consider the broad jump, in which the person jumps forward towards Link 1. The motion of the body parts produces very high speeds in the $+\mathbf{x}$ direction, moderate speeds in the $\pm\mathbf{z}$ direction, and negligible speeds in the $\mathbf{y}$ direction.

As discussed in Sec. 4.2.1, the instantaneous frequencies of the baseband received signal carry the information on the velocity components of different body parts. Thus, we utilize the frequency content of the received signals at the 3 links to construct informative features that can describe the person's motion characteristics during the gym activity. Towards this goal, we carry out time-frequency analysis of the simulated wireless signals, in order to extract features and train a classifier, as we discuss in the following part.

Time-Frequency Analysis: We estimate the frequency content of the signal via performing time-frequency analysis, a common method in harmonic analysis, which has also been used in RF sensing [1, 29]. More specifically, given a received signal in the time domain, we utilize Short-Time Fourier Transform (STFT) with windows of size 0.4 sec and a window overlap of 0.35 sec to generate a *spectrogram*, which contains the frequency

content of the signal (in the range of $[1, 100]$ Hz) as a function of time. In order to enhance the quality of the spectrogram, we carry out a denoising process, as follows. We first zero out all the spectrogram values that are below a noise floor of 0.01. Then, we binarize the spectrogram with a threshold of 0.01 and extract all the connected components from the 2D binary plot. Regions that correspond to very small components are zeroed out, since they are less likely to be the result of continuous human movements.

Fig. 4.5 (b) shows a sample spectrogram of the simulated signal at Link 3, for one repetition of stiff-leg deadlift extracted from the video of Fig. 4.5 (a). For instance, when the person moves into a bend-forward position as well as when she moves back to the initial position, the vertical speed of her body is clearly captured by the spectrogram of Link 3, in the first half (from 0 to 1 sec) and the second half (from 1 to 2 sec) of the repetition, respectively. In particular, in the video, the motion of going back to the standing position is faster than that of bending down. This is captured by the spectrogram where the frequency components are higher in the second half of the repetition, as compared to the first half. Note that the spectrogram (non-DC part) captures the frequency components of the motion and does not depend on the actual received signal strength as long as it is above the noise floor.

Features: Given the simulated signals of one repetition from a gym activity video, we first generate the corresponding spectrograms for the 3 links. We then extract several informative features from the spectrogram of a repetition in order to train a classifier.[5] Our proposed features are mainly based on the quantiles of the spectrograms as a function of time. More specifically, the q quantile of a spectrogram $S(f, t)$ is given as follows, as

[5]For a video that contains multiple repetitions of the same gym activity, we first temporally segment the video to extract each repetition (see Sec. 4.2.2) and treat each repetition as an individual training data point.

a function of time:

$$Q(t;q) = \min\left\{ f_q : \frac{\sum_{i=1}^{f_q} S(i,t)}{\sum_{i=1}^{f_{\max}} S(i,t)} \geq q \right\}, \tag{4.6}$$

where $Q(t;q)$ is the q quantile of the spectrogram as a function of time and $f_{\max}$ is the upper bound of the frequency in the spectrogram (i.e., 100 Hz in our case).

The quantiles as a function of time capture the temporal variations of the spectrogram which capture the time-varying speeds of the body parts, while staying robust to noise. In this study, we use the 0.5 and 0.7 quantiles for each spectrogram, which capture the median speed and the higher-speed components of the motion over time, respectively. Fig. 4.5 (c) shows the 0.5 quantile for the sample spectrogram of Fig. 4.5 (b).

Consider one repetition of an activity. We then calculate the histograms of the 0.5 quantile and the 0.7 quantile, respectively, within the repetition, for each spectrogram of each link. Each histogram is a 5-dimensional vector that contains the respective numbers of points with quantile value in the following intervals: $[1, 10]$ Hz, $(10, 20]$ Hz, $(20, 30]$ Hz, $(30, 40]$ Hz, and $(40, 100]$ Hz. These histograms efficiently capture the distributions of the quantile values in each spectrogram of the repetition. In order to capture any possible temporal asymmetry within one repetition of an activity, we calculate the difference between the maximum value of the quantile values in the first and the second halves of the repetition, and use a binary number to indicate whether the difference is larger than 10 Hz, for each quantile curves of Link 3. Finally, the time duration of the repetition is used as the last feature. This amounts to a total of 33 features for each repetition of an activity. Overall, these features capture various time and frequency attributes of the motion, which are useful for classifying the activities.

Since in the operation phase, the person may not exactly stand at the center (see Fig. 4.3), we further perturb the extracted meshes as follows, in order to augment the simulated dataset. More specifically, to generate a perturbed dataset, we draw two

numbers uniformly distributed in $[-0.1, 0.1]$ m, which we then use to shift the x and y positions of the 3D human mesh, respectively, in the GCS. We perturb each mesh 10 times in this manner. Overall, we have a total of 1878 simulated feature vectors to be used for training the classifier. As the number of repetitions differs for each activity class, we apply oversampling [73] to balance the training data.

Training a Classifier: We then train a linear classifier using the feature vectors of our simulated gym activity RF dataset. In the operation phase, the trained classifier then takes as input the features of one individual repetition from a measured RF signal and outputs a predicted probability distribution over the 10 activity classes via the softmax operation. The class that corresponds to the highest predicted probability is taken as the classification decision for the input data sample. Given an activity period that possibly contains more than one repetition of the same activity, we can also fuse the predictions of the individual repetitions, in order to achieve a more accurate overall classification. More specifically, we can average the predicted probability distributions of all the repetitions within the same activity period. The activity class corresponding to the highest predicted probability in the aggregated distribution then serves as the classification decision for the activity period.

4.2.3 Test Experiments with Real WiFi

We have conducted a number of test experiments to collect real WiFi measurements and evaluate the performance of our gym activity classification system that is trained only with online video data. In this section, we then discuss our WiFi experimental setup, and the test data collection and processing.

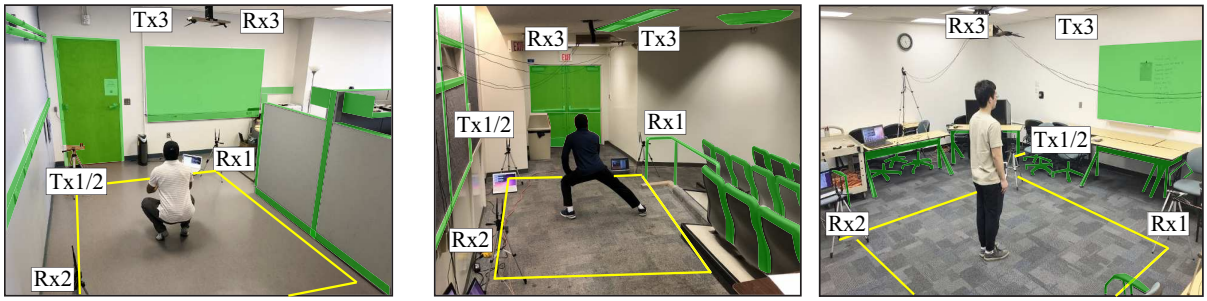


Figure 4.6: Our experimental setup in 3 different areas to test our gym activity classification system. Link 1 consists of Tx1/2 and Rx1, Link 2 consists of Tx1/2 and Rx2, and Link 3 consists of Tx3 and Rx3. Area 1 is in a lab, Area 2 is in the back of a classroom, and Area 3 is in a conference room. As can be seen, the areas are cluttered with a variety of objects. For instance, all metallic objects, which can be strong reflectors, are highlighted in green. See the color pdf to best view this figure.

Experimental Setup and Data Processing

The test experiments are conducted in a total of 3 test areas, which are shown in Fig. 4.6. These test areas represent several real-world environments with a variety of area sizes, geometry, and clutter. More specifically, Area 1 is located in a lab, Area 2 is located in the back of a classroom, while Area 3 is located in a conference room. Moreover, each test area contains several metallic objects of various sizes (marked with green), e.g., white board, desks, and chairs, which makes the test areas similar to the metal-rich environment in a gym. Clutter, however, does not impact the performance of our approach as the impact of static objects appear in the DC term and is removed, as discussed earlier.

Each Tx/Rx in the test area consists of $N_T = 2$ and $N_R = 2$ antennas connected to a laptop with Intel 5300 Network Interface Card. The shared transmitter for Link 1 and 2 transmits 400 WiFi packets per second on WiFi channel 36 ($f_c = 5.18$ GHz), while that of Link 3 transmits WiFi packets with the same rate on WiFi channel 44 ($f_c = 5.22$ GHz). We use CSItool [70] on each of the receivers to log the squared

magnitude of the Channel State Information (CSI) of the received packets. Each receiver logs a total of $N_T \times N_R \times 30$ subcarriers $= 120$ data streams, which we first denoise using Principal Component Analysis as described in [29]. More specifically, we use the first 10 principal components of the data streams and generate the received signal's spectrogram using the method of Chapter 2, where both STFT and Hermite functions are used to generate high-quality spectrograms. Note that the static multipath from the static objects in the environment appears at DC in the spectrogram, and can easily be removed by subtracting the mean of the signal before generating the spectrogram. We then denoise the spectrograms using the same denoising scheme as described in Sec. 4.2.2. However, here we adaptively estimate the noise floor of the spectrogram as the 99th percentile of the spectrogram values above 70 Hz. We assume that the frequency range above 70 Hz has no informative reflections from the human body, since it corresponds to speeds above 2 m/s.

WiFi Test Data Collection

For the WiFi test experiments, we have recruited a total of 10 subjects to participate in our test experiments, including 8 males and 2 females. In each area, each subject is asked to perform the 10 gym activities. For each activity, we collect the WiFi measurements of each subject for 45 seconds, to which we refer as an activity period. During each activity period, the subject performs multiple repetitions of this activity. We then temporally segment the WiFi measurement of each activity period to extract the time intervals of the individual repetitions, based on the brief resting periods between two consecutive repetitions.

Overall, we have a total of 1543 repetitions for the 10 gym activities, or equivalently, a total of 300 activity periods (100 activity periods per area), from the 10 subjects in the 3 areas. The activity periods each contain an average of 5.1 repetitions. More specifically,

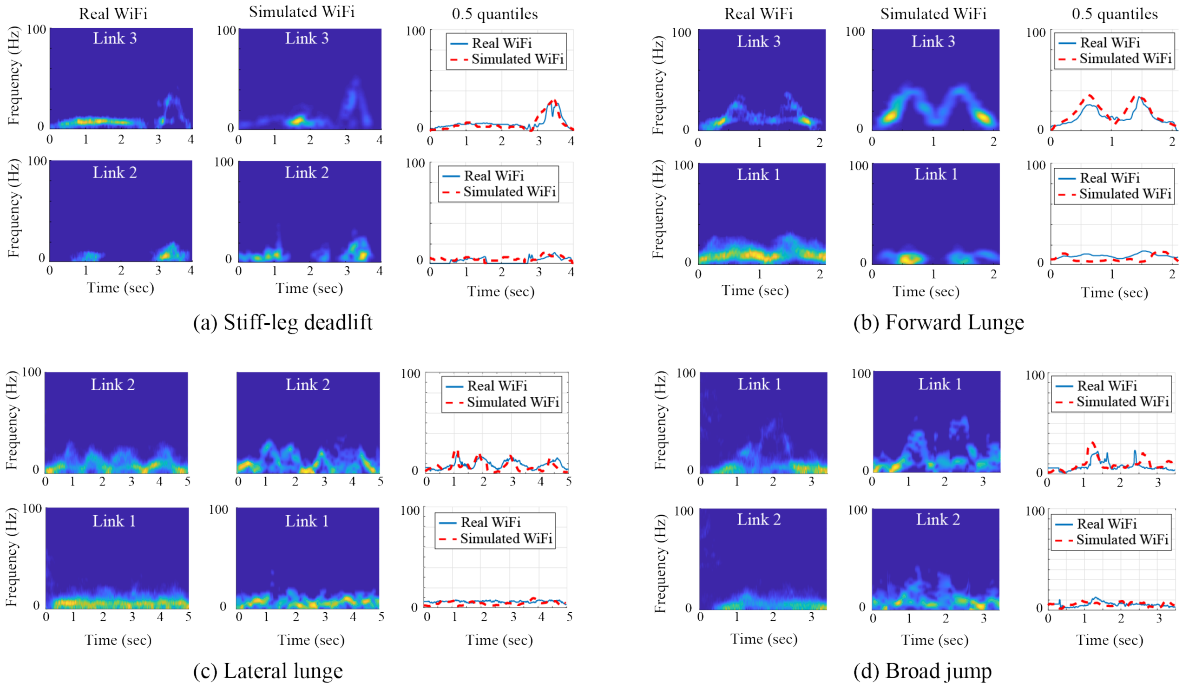


Figure 4.7: Comparison between the real and simulated WiFi spectrograms on two links for four exercises: (a) Stiff-leg deadlift, (b) Forward Lunge, (c) Lateral lunge, and (d) Broad jump. For each exercise, left figure is the real WiFi spectrogram, middle figure is simulated WiFi spectrogram, and right figure is the extracted 0.5-quantiles from the two spectrograms. It can be seen that the real and simulated spectrograms are visually similar, and the 0.5-quantiles confirm their similarity. Note that for each exercise, we show the two links that have the best (top) and the worst (bottom) matches between the real and simulated WiFi spectrograms.

we have 523 individual repetitions in area 1, 517 in area 2, and 503 in area 3.

4.2.4 Performance Evaluation

In this section, we evaluate the performance of our proposed approach for training the WiFi gym activity classifier using only video data and no RF data. We first analyze the similarity between the simulated and real WiFi signals, in order to validate our generated training data. Then, we extensively evaluate the classification performance of our trained WiFi sensing system with real WiFi test data.

Similarity between Simulated and Real Data

In order to present a proper assessment of the similarity between the simulated WiFi signal and the real one, we collect the WiFi measurements of a person performing two activities (stiff-leg deadlift and forward lunge), while recording a video of the scene at the same time. We then analyze the similarity of the simulated WiFi signal and the real one via spectrogram analysis, since a spectrogram can effectively capture the motion of different gym activities, as discussed in Sec. 4.2.2. Note that the video recording and WiFi measurements of this experiment are collected solely for the purpose of showing the similarity between the simulated data and the real one, and neither of them was used in the training set or in the test set of our system.

Fig. 4.7 shows the comparison between the real and simulated WiFi spectrograms for four sample exercises. For each activity, we show the two links that have the best (top) and the worst (bottom) matches between the real and simulated WiFi spectrograms. Fig. 4.7 (a) shows the spectrograms of the measured WiFi data on Link 2 and Link 3 for one repetition of the stiff-leg deadlift exercise (left column), as well as the video-based simulated WiFi data on these 2 links for the same exercise (middle column). It can be seen that the spectrograms based on simulating from the video properly capture the motion patterns, and match the WiFi spectrograms well. The figure also shows the 0.5 quantile curves (one of the features we shall use) of both the real WiFi spectrogram and the simulated one (right column), showing a good match between the two. Similarly, a good match can be seen between the real WiFi spectrograms and the simulated ones of the forward lunge, lateral lunge, and broad jump activities. The average cosine similarity between the 0.5 quantile curves of the simulated and real spectrograms of all four activities is 0.88, while two identical curves have a cosine similarity of 1. Finally, the spectrograms of these exercises reveal the unique patterns of each exercise in terms of the

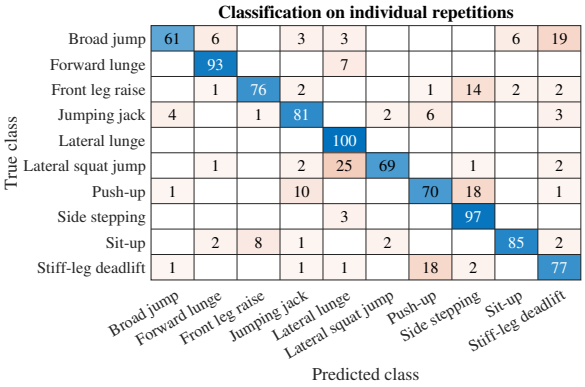


Figure 4.8: Confusion matrix of classifying the 10 gym activities with WiFi, based on individual repetitions of the activities.

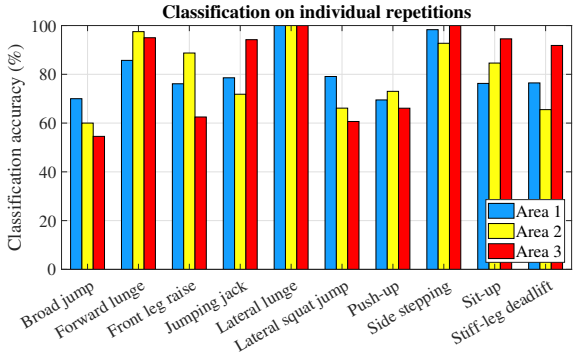


Figure 4.9: Classification accuracy for each activity class in each test area, when using individual repetitions.

frequency content of the spectrograms, or, equivalently, the speed profile of the person's body.

Overall, Fig. 4.7 shows the power of our proposed approach in generating simulated RF data that closely resemble the real one, and highlights its potential in eliminating the need for the collection of real RF measurements when training RF sensing systems.

Classification Performance

In this part, we evaluate the performance of our WiFi-based gym activity classifier, which is trained with only video data. We first present the results for the case where the classifier is tested with individual activity repetitions. In other words, multiple repetitions of the same activity by the same person are treated as independent test cases in this setting. We then evaluate the classifier for the case where it jointly uses all the repetitions done by the same person during an activity period to classify his/her activity (by fusing their corresponding decisions). As expected, the second case would perform better since the data of a few repetitions is used for classification. The first case, however, is important as it establishes a lower bound on the performance for the case when the person only does one repetition of an exercise. As we shall see, we can still classify the

activities well, even with only one repetition.

Classification on Individual Repetitions: In this setting, our classifier achieves an average classification accuracy of 81%, over all the test areas (a random selection would have resulted in 10% accuracy). Fig. 4.8 shows the confusion matrix that corresponds to the classification performance of our trained system on individual repetitions. The diagonal entries indicate the classification accuracy for each activity (in %) and the off-diagonal entries indicate the percentages of misclassifications to that corresponding class. Overall, it can be seen that our system can classify all the activities pretty well. In particular, activities such as forward lunge, lateral lunge, and side stepping are recognized very well, with classification accuracies above 90%. On the other hand, some other activities, such as lateral squat jump, are classified with a lower accuracy due to the inherent similarity with another activity. For instance, lateral squat jump is very similar to lateral lunge, as they both require lateral and vertical motion of the body.

Next, we show the classification accuracy as a function of the areas in Fig. 4.9. It can be seen that the respective accuracies in the three areas are very similar to each other. This indicates that our video-trained WiFi sensing system is not sensitive to the multipath effects from static objects in real WiFi environments, due to the fact that we utilize motion-driven features that capture only the person's speed profile, as discussed in Sec. 4.2.2.

Classification on Activity Periods: We next show the performance when a small number of repetitions of an activity period are used to classify that activity. During the test, each person performs each activity for 45 seconds in each area. Thus, depending on the speed of the person, there may be more than one repetition in an activity period (the average is 5.1 repetitions per activity period). For such a case, our classifier fuses the predictions of the individual repetitions in order to improve the classification quality,

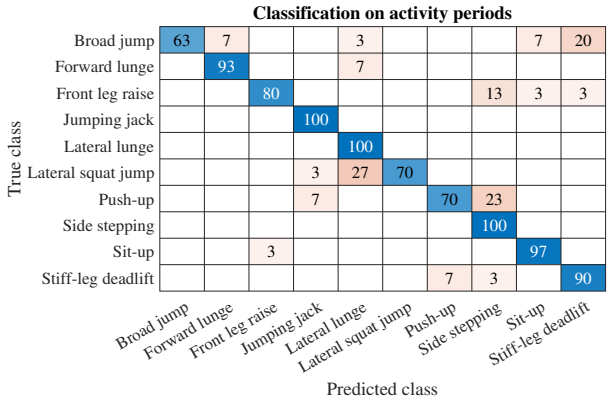


Figure 4.10: Confusion matrix of classifying the 10 gym activities with WiFi, based on activity periods containing an average of 5.1 repetitions each.

as discussed in Sec. 4.2.2. Fig. 4.10 shows the confusion matrix for this case. The overall average classification accuracy improves to 86%, as compared to the accuracy of 81% on individual repetitions. In particular, jumping jack and side stepping are now classified correctly 100% of the time, as compared to 81% and 97% in the repetition-based setting. This case also has a similar performance across the 3 areas.

Classification Error Analysis: In this part, we perform an in-depth analysis on the classification errors. Fig. 4.11 shows the spectrograms for a sample activity pair that is confusing for the classifier (push-up and side stepping), as indicated in the confusion matrices in Figs. 4.8 and 4.10. More specifically, push-ups are classified as side stepping 23% of the time according to Fig. 4.10. It can be seen that the frequency distribution in each link is very similar for the two activities, due to the inherent motion similarity between the two, which is the main source of classification error. For instance, since neither of them involve highly-dynamic motion, the corresponding frequencies captured in all the links are low for both activities, with similar spectrogram patterns.

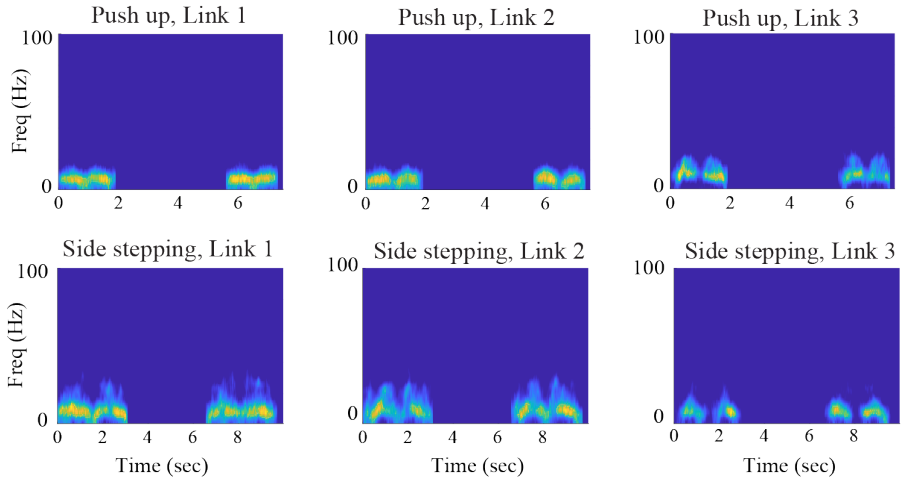


Figure 4.11: WiFi spectrograms of two activities: push up and side stepping, which are the very confusing for the classifier (push-ups are classified as side stepping 23% of the time as can be seen in Fig. 4.10). Two repetitions of the exercise are shown in each spectrogram. It can be seen that the frequency distribution is similar for both activities, which makes it more challenging for the classifier to differentiate them.

Robustness to Metallic Objects in the Environment

Our test areas of Fig. 4.6 contain various metallic objects (comparable to human heights) around the person to resemble the gym environment. In our proposed system, we utilize non-DC parts of the spectrograms to capture the motion information, which are insensitive to the static scatterers in the environment which appear at DC, even if they are highly reflective. In order to further test the robustness of our pipeline to highly reflective clutter, we have created an even more metal-rich environment by including additional large highly-reflective objects (e.g., multiple metallic drawer cabinets and highly-reflective shielding material sheets) in one of our test areas, as shown in Fig. 4.12. One subject then performs different exercises in both the original setting of this area (top row) and the new metal-heavy setting (bottom row). Fig. 4.12 also shows the spectrograms of the received WiFi signals in both settings, for four sample exercises. It can be seen that the spectrograms of the received signals are almost identical, with or without the large metallic objects in the area. This result validates our signal model of Eq. 4.4 and indicates

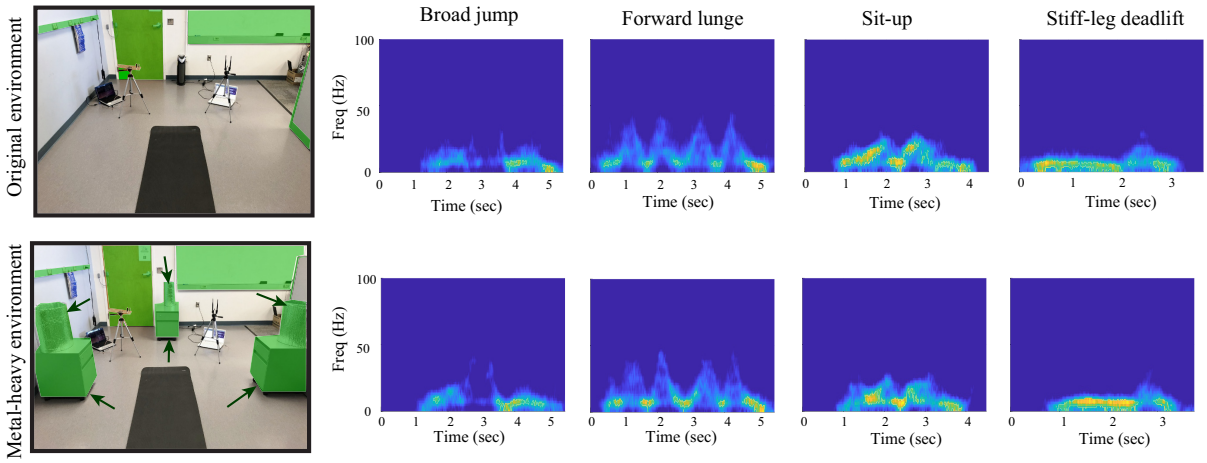


Figure 4.12: Spectrograms of the collected WiFi data on Link 1 for a person performing 4 sample exercises in both the original setting (top row) and a metal-heavy setting (bottom row) of Area 1. Metallic objects are highlighted in green. The arrows point to the added metallic objects in the metal-heavy setting of the bottom figure.

that our proposed pipeline is not affected by the static objects in the environment, even if they are highly reflective.

Overall, our results show that we have, for the first time, successfully trained an RF sensing system, without collecting any prior RF training measurements. Moreover, in terms of RF-based gym activity classification, our proposed approach has not only enabled it with no prior RF training measurements, but has also enabled the first reflection-based system, whereas all the existing methods rely on the person to cross/block the line-of-sight path (i.e., the direct path between the Tx and Rx) while performing the exercises.

4.3 Discussions

In this section, we discuss a few more aspects related to our proposed framework and the case study.

Generalization of the Video-Based Training Approach to Other Applications:

In this chapter, we proposed a general approach that can train RF sensing systems without any RF training data, and by using the vast available online videos. While we showcased the performance of this approach in the context of gym activity recognition, showing how we can achieve a high accuracy with zero RF training data, the proposed methodology is applicable to many other RF sensing applications, scenarios, and setups. As part of the future work, we envision that this approach can be used to train other RF sensing systems to recognize other activities, gestures, and in general other situations that involve motion of body parts. It can further be used for analysis purposes, for instance to understand the optimal RF setup/amount of needed resources for a particular application, to understand the differentiability of different activities, or to understand the limitations of sensing with a certain setup or at a particular frequency, all without the need to collect any RF data. Overall, the proposed approach is scalable and general, and can thus enable new work in the area of RF sensing.

Further Discussions on the Sensing Setup: In the considered sensing setup for our gym activity classification study (Fig. 4.3), we assumed that the person is at the center of the coordinate system and facing the positive **x** direction. In order to set the coordinate system of the sensing setup in that manner, the location and orientation of the person are assumed to be known. This is a realistic assumption since there is a great body of work on localization and tracking with RF signals, e.g., [96, 97, 98], that can be utilized to first estimate the location and orientation of the person.

Furthermore, we assumed that the 3 WiFi links are placed such that $\vec{\mathbf{u}}_{\ell_1}$ is parallel to the **x**-axis, $\vec{\mathbf{u}}_{\ell_2}$ is parallel to the **y**-axis, and $\vec{\mathbf{u}}_{\ell_3}$ is parallel to the **z**-axis (see Fig. 4.3). As such, in this configuration, each link directly captures the information about one of the three-dimensional components of the velocity vectors of different body parts (as discussed in Sec. 4.2.1). For any general transceiver locations that are not similar to

Fig. 4.3, the motion information across the three dimensions can become coupled in the measurements. Then, a simple linear system equation can be solved to directly extract the motion components across the three dimensions, similar to what is done in [99]. Once these are extracted, our trained pipeline based on the configuration of Fig. 4.3 can be used for classification. In summary, the configuration of Fig. 4.3 can be used as a base since it directly measures the motion across the three dimensions while other configurations can be translated to it.

Sensing Multiple People: In the gym activity classification study, we assumed that there is only one person doing the exercise, with little movements from other people nearby (other people were present but not moving much). In the case where there are other people simultaneously moving nearby (e.g., performing exercises), the received signal will contain the motion information of the person of interest as well as those nearby. In such scenarios, one can then use multiple antennas at each transceiver to create a small antenna array and separate the impact of multiple people on the received signal by beamforming towards each person [77]. The reflected signals off of different people can also be separated in other domains, e.g., Time-of-flight, Angle-of-Departure [100].

Chapter 5

Nocturnal Seizure Detection Using Off-the-Shelf WiFi

In this chapter, we propose a novel healthcare application for sensing with WiFi signals. More specifically, we propose a new mathematical foundation that enables WiFi signals, for the first time, to detect nocturnal seizures in epilepsy patients. Using WiFi signals for this task has several advantages when compared to traditional seizure detection techniques, such as being contactless, cheap, fast, and robust. We start this chapter by mathematically analyzing the received signal during different kinds of nocturnal motions: breathing, normal sleep movements, and seizures.

Remark 5.1 *In this chapter, we use the term seizure to refer to tonic-clonic seizure, which is a type of seizures that happens more frequently during sleep [101]. Tonic-clonic seizures are characterized by a tonic phase, in which the body muscles stiffen for a few seconds, followed by a clonic phase, in which the body muscles rapidly and rhythmically jerk for 1-3 minutes [102]. We also use the term normal sleep events to refer to normal non-breathing body movements during sleep, such as pose adjustments, stretching, scratching, coughing, sneezing, jerking, and others.*

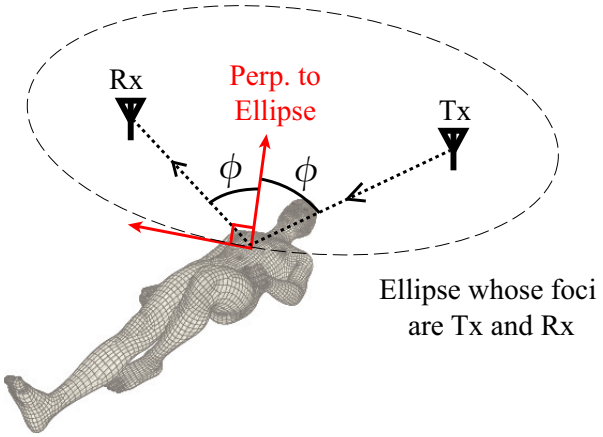


Figure 5.1: Illustration of the application scenario. A pair of WiFi transceivers collect WiFi CSI measurements while a person is sleeping, in order to analyze their sleep motions and detect if they are having a seizure. Note that our design does not assume or require that the person lies on their back, and they can be in any pose/orientation. Furthermore, the TX/RX can be in any configuration as well.

5.1 Signal Model

In this section, we develop a mathematical model for the received WiFi CSI in a general setting, an example of which is shown in Fig. 5.1. More specifically, a person is lying down on a bed in any generic pose while a WiFi transmitter (Tx) emits wireless signals that are reflected off of the person's body and received by a WiFi receiver (Rx). We first derive closed-form expressions for the WiFi CSI squared magnitude and the WiFi CSI phase difference signals, during a generic motion pattern of the body, in this part. In Sec. 5.2, we then use this model to provide a new and rigorous mathematical analysis of the spectral content of the WiFi signals during specific kinds of sleep motions relevant to this work, i.e., breathing, seizures, and normal sleep movements.

Let $c(t)$ denote the complex baseband received signal at the Rx, which can be decomposed into the direct path from the Tx to the Rx, and the reflected path off of the

person's moving body. More specifically, $c(t)$ can be written as [67]

$$c(t) = \underbrace{\alpha_d e^{j\mu_d}}_{\text{direct path}} + \underbrace{\alpha_r e^{j\left(\mu_r + \frac{2\pi}{\lambda}\psi \int v(t)dt\right)}}_{\text{reflected path}}, \tag{5.1}$$

where α_d and μ_d are the amplitude and phase of the direct path from the Tx to the Rx, α_r is the amplitude of the reflected path arriving at the Rx, μ_r is the phase of the reflected path at time $t = 0$, $\psi = 2\cos(\phi)$ is a scale parameter that depends on the location of the bed/person with respect to the Tx and Rx. Consider the ellipse whose foci are the Tx and Rx, which passes through the person's body, ϕ is then the angle between the line connecting the person to the Tx (or Rx) and the perpendicular line to this ellipse at the point that it passes through the person's body (see Fig. 5.1). $v(t)$ is the instantaneous speed component of the body motion along the perpendicular line to the ellipse, and λ is the wavelength.

Note that the value of ψ depends on the scene configuration, i.e., the relative location of the bed with respect to the Tx and Rx, and does not depend on the person's posture and orientation while sleeping. In other words, if the width of the bed is small as compared to the Tx-Rx distance, or the person does not drastically move from one side of the bed to the other (which is common in practice), the sleeping person's general location with respect to the Tx and the Rx does not drastically change, and hence, ψ can be taken as a constant and can be calculated only once upon Tx-Rx placement. The person can change their pose/orientation several times, but those movements will not affect the value of ψ.

For simplicity of notation, we define $\beta = \frac{2\pi\psi}{\lambda}$. Hence, the phase of the reflected path at the Rx becomes $\mu_r + \beta \int v(t)dt$. Next, we derive closed-form expressions for the squared magnitude and phase of $c(t)$ to understand the information they carry about the body's motion.

Remark 5.2 *The static multipath in the environment does not affect this analysis since all the static multipath can be integrated into the first term of Eq. 5.1. This indicates that the performance of the system is agnostic to the deployment environment. This observation will be further validated by our extensive experiments in several different locations and real-world scenarios, as we shall see in Sec. 5.5.*

Squared Magnitude of $c(t)$: The squared magnitude of $c(t)$ can be written, after a straightforward derivation, as follows,

$$|c(t)|^2 = c(t)c^*(t) = \alpha_d^2 + \alpha_r^2 + A_m \cos\left(\beta \int v(t)dt + \Delta\mu_m\right), \tag{5.2}$$

where $A_m = 2\alpha_d\alpha_r$, and $\Delta\mu_m = \mu_r - \mu_d$ is the difference between the initial phase of the reflected path and the phase of the direct path. Since the DC component of $|c(t)|^2$ does not carry any information about the motion of the body, we subtract the DC term (which can be easily implemented in practice) to have the following,

$$s_m(t) = A_m \cos\left(\beta \int v(t)dt + \Delta\mu_m\right). \tag{5.3}$$

For the ease of discussion, we then refer to $s_m(t)$ as the squared magnitude signal in the rest of the chapter.

Phase of $c(t)$: Without loss of generality, we analyze the phase of the scaled signal $c'(t) = e^{-j\mu_d}c(t)/\alpha_d$. This scaling shifts the phase of $c(t)$ by a constant amount, preserving the time-varying behavior of the phase of $c(t)$ which carries the motion information of the body. Let $\angle c'(t)$ be the phase of $c'(t)$. It is easy to confirm that

$$\angle c'(t) = \tan^{-1}\left(\frac{\frac{\alpha_r}{\alpha_d}\sin\left(\beta\int v(t)dt + \Delta\mu_m\right)}{1 + \frac{\alpha_r}{\alpha_d}\cos\left(\beta\int v(t)dt + \Delta\mu_m\right)}\right). \tag{5.4}$$

Due to its longer length and the reflection loss at the body, we can assume that the

amplitude of the reflected path is much less than that of the direct path, i.e. $\frac{\alpha_r}{\alpha_d} \ll 1$. In such a case, $\angle c'(t)$ can be approximated as

$$\begin{aligned}
\angle c'(t) &\approx \tan(\angle c'(t)) \\
&\approx \frac{\alpha_r}{\alpha_d} \sin\left(\beta \int v(t)dt + \Delta\mu_m\right) \left(1 - \frac{\alpha_r}{\alpha_d} \cos\left(\beta \int v(t)dt + \Delta\mu_m\right)\right) \\
&= \frac{\alpha_r}{\alpha_d} \sin\left(\beta \int v(t)dt + \Delta\mu_m\right) - \frac{\alpha_r^2}{2\alpha_d^2} \sin\left(2\beta \int v(t)dt + 2\Delta\mu_m\right) \\
&\approx \frac{\alpha_r}{\alpha_d} \sin\left(\beta \int v(t)dt + \Delta\mu_m\right), \qquad (5.5)
\end{aligned}$$

where the first order Taylor approximation $(1+x)^{-1} \approx 1 - x$ for $x \ll 1$ is used in the second line to derive Eq. 5.5, since $\frac{\alpha_r}{\alpha_d} \ll 1$.

In practice, the phase measurements on off-the-shelf WiFi devices are corrupted by multiple sources of error, such as Carrier Frequency Offset (CFO) and Sampling Time Offset (STO), rendering these phase measurements unreliable [17]. However, since different antennas of the same WiFi card share the same oscillator, those errors are common to all the antennas of the same card, and as such, the phase difference between two antennas of the same card carries stable phase information, as has been used in the literature. In this chapter, we also rely on the phase difference between the antennas of one receiver WiFi card. Let $\angle c_i(t)$ be the phase of the CSI at the i-th antenna of the Rx. The phase difference between the i-th and j-th receiver antennas can then be written as

$$s_p(t) = \angle c_i(t) - \angle c_j(t) = A_p \cos(\beta \int v(t)dt + \Delta\mu_p), \qquad (5.6)$$

where $A_p = 2(\alpha_r/\alpha_d) \sin(0.5(\Delta\mu_{m,i} - \Delta\mu_{m,j}))$, $\Delta\mu_p = 0.5(\Delta\mu_{m,i} + \Delta\mu_{m,j})$, and $\Delta\mu_{m,i}$ and $\Delta\mu_{m,j}$ are the values of $\Delta\mu_m$ at the i-th and j-th receiver antennas, respectively.

Eq. 5.6 shows that as long as $\Delta\mu_{m,i} \neq \Delta\mu_{m,j}$ (which depends on the direct and

reflected path lengths to the receiver antennas as well as the wavelength), the phase difference between the receiver antennas has a similar structure, in terms of the information it carries about the body movements, as the squared magnitude of the received signals (Eq. 5.3).

Body Acting as an FM Radio: Frequency Modulation (FM) is a classic analog transmission technique, introduced in 1902 [103], to ensure robust transmissions for radio applications. A typical FM transmitted signal will have the form $\cos(2\pi f_c t + k_{\text{f}} \int m(t)dt)$, where $m(t)$ is the signal of interest to be transmitted, f_c is the carrier frequency, and k_{f} is the modulation index constant. As can be seen, both the squared magnitude signal of Eq. 5.3 and the phase difference signal of Eq. 5.6 can be interpreted as FM signals, in which $v(t)$ is the modulating signal and $f_c = 0$. In other words, the moving body part (e.g., the chest) can be thought of as modulating the body motion into an FM signal that is then received by the WiFi receiver. This way of interpretation allows us to delve into the classic mathematical analysis of FM signals for our system design, as we shall see in the next section. However, one difference with a typical FM signal is the existence of the $\Delta\mu_m$ term in Eq. 5.3 (or $\Delta\mu_p$ in Eq. 5.6). We shall see the impact of such a term in the spectral analysis of the next section.

5.2 Spectral Analysis of the Received Signal

In this section, we analyze the received squared magnitude signal (or, equivalently, the phase difference signal) of Sec. 5.1, for different kinds of nocturnal body movements: breathing, seizures, and normal sleep events (e.g., posture shifts, moving limbs, etc.). More specifically, we develop our first major contribution: ***to mathematically characterize the spectral content/bandwidth of the received signal for each of the aforementioned three types of motions***. We shall see that, due to the different

body motion characteristics during a seizure as compared to normal sleep events, the spectral content of the received signals can be used to design a robust nocturnal seizure detection system, as we shall see in Sec. 5.3.

Let $y(t) = A\cos(\beta \int v(t)dt + \Delta\mu)$ represent a general form for either the squared magnitude signal of Eq. 5.3 or the phase difference signal of Eq. 5.6. First, assume that $v(t)$ is a sinusoidal signal of the form $v(t) = v_{\max}\cos(\omega_o t)$. This assumption applies to both the seizure and respiration cases. The following characterizes the Fourier response of $y(t)$.

Theorem 5.1 *Consider the signal $y(t) = A\cos(\beta \int v(t)dt + \Delta\mu)$ with a sinusoidal speed signal of $v(t) = v_{max}\cos(\omega_o t)$. The spectrum of this signal, i.e., its Fourier transform, can be written as follows,*

$$\begin{aligned} Y(f) = A\cos(\Delta\mu) \sum_{\substack{n \ even \\ n\geq 0}} J_n(\beta') \left(\delta(f - nf_o) + \delta(f + nf_o)\right) \\ + jA\sin(\Delta\mu) \sum_{\substack{n \ odd \\ n>0}} J_n(\beta') \left(\delta(f - nf_o) + \delta(f + nf_o)\right), \end{aligned} \tag{5.7}$$

where $J_n(.)$ is the n-th order Bessel function, $\beta' = \beta v_{max}/\omega_o$, $\delta(.)$ is the Dirac-Delta function, and $f_o = \omega_o/2\pi$ is the fundamental frequency of $v(t)$.

Proof: If $v(t) = v_{\max}\cos(\omega_o t)$, then $y(t)$ becomes

$$\begin{aligned} y(t) &= A\cos(\Delta\mu)\cos(\beta'\sin(\omega_o t)) - A\sin(\Delta\mu)\sin(\beta'\sin(\omega_o t)) \\ &= A\cos(\Delta\mu)\mathcal{R}\left\{e^{j\beta'\sin(\omega_o t)}\right\} - A\sin(\Delta\mu)\mathcal{I}\left\{e^{j\beta'\sin(\omega_o t)}\right\} \end{aligned} \tag{5.8}$$

where $\beta' = \beta v_{\max}/\omega_o$, $\mathcal{R}\{.\}$ is the real part of the argument, and $\mathcal{I}\{.\}$ is the imaginary part of the argument. The exponential term $e^{j\beta'\sin(\omega_o t)}$ is periodic with a period $2\pi/\omega_o$,

and can be expanded by its Fourier Series as [104]

$$e^{j\beta' \sin(\omega_o t)} = \sum_{n=-\infty}^{\infty} J_n(\beta') e^{jn\omega_o t}, \tag{5.9}$$

where $J_n(.)$ is the n-th order Bessel function. By substituting Eq. 5.9 into Eq. 5.8, we get

$$y(t) = \sum_{n=-\infty}^{\infty} A\, J_n(\beta') \left(\cos(\Delta\mu)\cos(n\omega_o t) - \sin(\Delta\mu)\sin(n\omega_o t)\right).$$

By making use of the fact that $J_{-n}(x) = (-1)^n J_n(x)$, $y(t)$ can be written as

$$y(t) = 2A\cos(\Delta\mu) \sum_{\substack{n \text{ even} \\ n\geq 0}} J_n(\beta')\cos(n\omega_o t) - 2A\sin(\Delta\mu) \sum_{\substack{n \text{ odd} \\ n>0}} J_n(\beta')\sin(n\omega_o t).$$

By taking the Fourier transform of $y(t)$, we get Eq. 5.7. ■

Theorem 5.1 states that the spectrum of $y(t)$ consists of an infinite number of deltas, located at the fundamental frequency of $v(t)$ and its harmonics. We next characterize the bandwidth of this signal. In order to do so, we need to find the frequency point after which the power of the subsequent delta functions has become negligible, as compared to the earlier terms.

Theorem 5.2 *The bandwidth of $y(t)$ can be characterized as follows, for $\beta' \geq 1$,*

$$BW|_{\beta'\geq 1} = (\beta' + 1) f_o = \psi v_{max}/\lambda + f_o,$$

where f_o is the fundamental frequency of $v(t)$. Moreover, for $\beta' < 1$, the bandwidth of

$y(t)$ is best characterized as follows,

$$BW|_{\beta'<1} = 2f_o.$$

Proof: It is well established in the literature that $J_n(\beta')$ is negligible for $n > \beta' + 1$ [104]. By applying this to Eq. 5.7, we can then estimate the bandwidth as follows: $BW = (\beta' + 1)f_o$ for $\beta' \geq 1$ since some even and odd terms are both present for $\beta' \geq 1$ and terms can be compared accordingly within each even and odd groups. When $\beta' < 1$, however, the previous result implies that the term $n = 1$ is the only dominating term in the spectrum of $y(t)$. However, due to the different scaling factors of the even and odd terms in Eq. 5.7, there could exist cases (e.g., small $\Delta\mu$) where the term corresponding to $n = 1$ is suppressed by the $\sin(\Delta\mu)$ factor. In such cases, even though $J_2(\beta')$ is small as compared to $J_1(\beta')$, $\cos(\Delta\mu)J_2(\beta')$ can be comparable or larger than $\sin(\Delta\mu)J_1(\beta')$. Higher order terms can always be neglected with respect to the first two terms. Hence, the bandwidth of $y(t)$ for the case of $\beta' < 1$ is $2f_o$. ■

Remark 5.3 *In his seminal paper of [105], J. Carson was the first to theoretically characterize the bandwidth of an FM signal and show that it can be larger than the bandwidth of the modulating signal. Carson has shown that his bandwidth rule is exact for sinusoidal modulating signals, but can be generalized to approximate the bandwidth for general non-sinusoidal modulating signals as well. As mentioned earlier, our received signal $y(t)$ has a close resemblance to an FM signal, except for the $\Delta\mu$ terms. As such, our bandwidth analysis has some resemblance to Carson's derivations except for the impact of $\Delta\mu$. Following a similar argument to Carson's, we will then also use Theorem 5.2 to approximate the bandwidth of $y(t)$ when $v(t)$ is a general non-sinusoidal signal in the next section. In such a case, f_o would denote the bandwidth of the signal $v(t)$.*

Theorem 5.2 states that the bandwidth of $y(t)$ depends on motion parameters such

as $v_{\max}$ and f_o (or the bandwidth of $v(t)$ for non-sinusoidal signals). We next utilize Theorem 5.2 to estimate the bandwidth of the WiFi CSI [1] during three specific kinds of sleep-related motions: breathing, seizure, and normal sleep movements.

5.2.1 CSI Bandwidth During Breathing

A sleeping person's chest volume expands and shrinks during the inhalation and exhalation phases of respiration. It is established in the literature that the instantaneous chest speed, i.e., $v(t)$ of Sec. 5.1, can be approximated by a sinusoid of frequency $f_{o,\text{br}}$, where $f_{o,\text{br}}$ is the number of breathing cycles per second [106]. As such, Eq. 5.7 can describe the spectrum of the WiFi signal during breathing.

In order to characterize the bandwidth for the case of normal breathing, we need to estimate $\beta'_{\text{br}} = \frac{\psi v_{\max,\text{br}}}{\lambda f_{o,\text{br}}}$, where $f_{o,\text{br}}$ is the breathing rate of the person, which is typically in the range of 0.2 to 0.3 Hz [107]. By integrating $v(t)$, it can be easily confirmed that $v_{\max,\text{br}}/2\pi f_{o,\text{br}}$ is equal to the maximum chest displacement during respiration, which has been reported in the literature to be around 5 mm [48]. This results in $\beta'_{\text{br}} \approx 0.55$ when using WiFi channel 48, which has a carrier frequency $f_c = 5.24$ GHz, and $\psi = 1$.[2] By using Theorem 5.2, we can then estimate the bandwidth of the received WiFi CSI during normal breathing as $BW_{\text{br}} = 2f_{o,\text{br}}$. Note that if the maximum chest displacement is not along the perpendicular line to the ellipse whose foci are the Tx and Rx (see Fig. 5.1), e.g., if the person is in a different pose, the chest speed will have a smaller velocity component along that line (i.e. smaller $v_{\max,\text{br}}$). In such a case, the value of β' will be even smaller and thus still less than 0.55. Thus, according to Theorem 5.2, the bandwidth

[1] Henceforth, the bandwidth of WiFi CSI means either the squared magnitude or phase difference, since they both have the same generic form $y(t)$.

[2] In our experiments, we use WiFi channel 48 ($f_c = 5.24$ GHz) in a setup in which $\psi \approx 1$ (see Sec. 5.4 for the detailed scene configuration). Extension to different values of ψ is straightforward, as we shall discuss in Sec. 5.5.2 where we show experiments with different ψs. Hence, we set $\lambda = 5.72$ cm and $\psi = 1$ for our numerical calculations in the rest of the chapter up to Sec. 5.5.2.

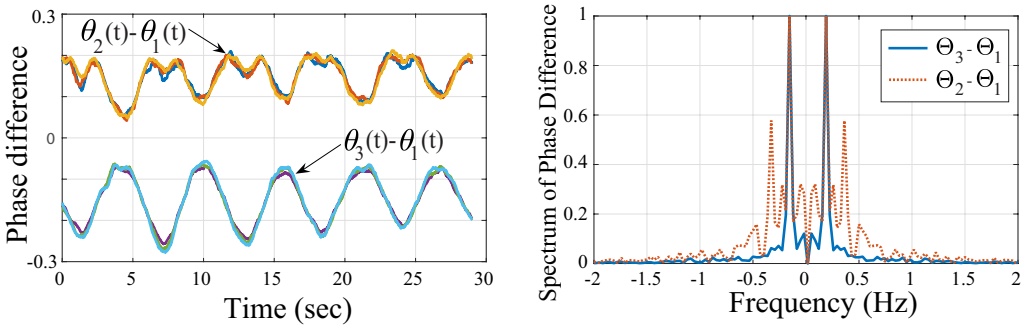


Figure 5.2: (Left) CSI phase difference at multiple subcarriers of the Rx and (right) the spectra of the CSI phase difference signals showing different spectral content for different antennas.

still remains $BW_{\text{br}} = 2f_{o,\text{br}}$ for all the cases.

It is worth noting that the previous literature on breathing monitoring using WiFi signals (either magnitude [108, 16, 106] or phase difference [49]) assume that the received WiFi signal rises and falls with the same frequency as the rise and fall (inhalation and exhalation) of the chest during the breathing process. Hence, they assume that the received signal has the same spectral content/bandwidth as that of the physical chest motion (i.e. they take BW_{br} to be $f_{o,\text{br}}$). However, Theorem 5.2 shows that the received WiFi signals can have a spectral content that is different from the physical breathing rate, depending on the value of $\Delta\mu$, with a maximum bandwidth of $2f_{o,\text{br}}$. To see this in effect, Fig. 5.2 shows the phase difference of the measured WiFi signals between the Rx antennas in a sample experiment, where a person was breathing with a frequency of $f_{o,\text{br}} \approx 0.18$ Hz. It can be seen that while the measured phase difference between antennas 3 and 1 of the Rx has a sinusoid-like pattern similar to that of the breathing motion, the phase difference between antennas 2 and 1 (due to having a different $\Delta\mu$) is experiencing a different pattern, which has a strong frequency component at $2f_{o,\text{br}}$.

5.2.2 CSI Bandwidth During Seizures

As described earlier, a tonic-clonic seizure consists of a *tonic* phase, in which the body muscles stiffen for a few seconds, immediately followed by a *clonic* phase, which is a strong, fast, and repeated stiffening and relaxing of the body muscles that can last for 1 to 3 minutes [102]. Several medical studies have been conducted to analyze body motion during a tonic-clonic seizure through data obtained by accelerometry. These studies have found that during the clonic phase of a tonic-clonic seizure, the body muscles rhythmically stiffen and relax with a frequency $f_{o,\text{sz}}$ between 1.5 and 5 Hz [109, 110, 52], thus making a sinusoid a good approximation for $v(t)$. Therefore, Eq. 5.7 also characterizes the frequency spectrum of the WiFi CSI during a seizure. In order to find the value of the parameter $v_{\max}$, and thus β', we have looked extensively into the medical literature on seizures. Several papers have found that the maximum acceleration, $a_{\max}$, of the body parts during a tonic-clonic seizure typically exceeds 15 m/s^2 [53, 54]. Since $v(t)$ is sinusoidal, then $v_{\max,\text{sz}} = \frac{a_{\max}}{2\pi f_{\text{sz}}}$, and a lower bound for the value of $v_{\max,\text{sz}}$ can be calculated as $v_{\max,\text{sz}} \geq \frac{15}{2\pi\times 5} = 0.48$ m/s.

Based on the aforementioned seizure motion parameters, one can estimate a lower bound for the bandwidth of the WiFi CSI during a seizure using Theorem 5.2 as $BW_{\text{sz}} = (\beta'_{\text{sz}} + 1)f_{\text{sz}} = \frac{\psi v_{\max,\text{sz}}}{\lambda} + f_{o,\text{sz}}$. More specifically, by using WiFi channel 48 and $\psi = 1$, $v_{\max,\text{sz}} = 0.48$ m/s and $f_{o,\text{sz}} = 1.5$ Hz, a lower bound for the bandwidth of the WiFi signal during the seizure is estimated as $BW_{\text{sz}} \geq 9.9$ Hz. Note that the aforementioned characterization of the CSI bandwidth during a seizure assumes that the motion of at least one body part is aligned with (or has a strong component on) the perpendicular line to the Tx-Rx ellipse of Fig. 5.1. This assumption is practical since the uncontrolled muscle jerks during the seizure result in the body parts moving randomly in all different directions. Moreover, it has been shown in the medical literature that a patient's body

posture can change to many different positions during a seizure [111]. Therefore, there will at least be one body part whose motion direction is aligned with the perpendicular line to the Tx/Rx ellipse.

It is worth stressing that the traditional assumption that the WiFi signal rises and falls with the same frequency of the body motion will result in a bandwidth estimation of $BW_{\text{sz}} = f_{o,\text{sz}}$, which is far off from the true bandwidth during a seizure.

5.2.3 CSI Bandwidth During Normal Sleep Events

We next delve into the medical literature on sleep motion analysis in order to characterize the parameters relevant for signal bandwidth characterization during normal sleep events, such as position adjustments and jerking in limbs, which people tend to make during different stages of sleep. It is found that these normal sleep events occur at an average rate of 3 events per hour [112], and can last for up to 15 seconds each [113]. Furthermore, other studies have performed time-frequency analysis of the accelerometry data of normal sleep and established that most of the power of normal sleep event signals (e.g. $v(t)$) is concentrated below $f_{o,\text{nm}} = 2$ Hz [114]. While $v(t)$ is non-sinusoidal, and no exact closed-form expression exists for the spectrum of the CSI signals for a general $v(t)$, Theorem 5.2 can still be used to approximate the bandwidth of the WiFi CSI, as discussed in Remark 5.3.

In order to calculate an upper bound for the WiFi CSI bandwidth in case of normal sleep events, we focus on wrist movements during sleep, which can have higher speeds due to its relatively lower mass as compared to other body parts. We utilize the online dataset published by the authors of [115] for the accelerometry data of 31 adults during their sleep, collected from wrist-worn Apple watches. By integrating this acceleration data over time, we get the instantaneous speeds of the wrist during sleep. We then use

	Breathing	**Seizure**	**Normal Event**
Motion Frequency (Hz)	$f_{o,\text{br}} =$ 0.2-0.3	$f_{o,\text{sz}} =$1.5-5	$f_{o,\text{nm}} =$ 2
$v_{\max}$ (m/s)	$\leq$0.01	$\geq$0.48	$\leq$0.33
BW of WiFi signal (Hz)	$BW_{\text{br}} = 2f_{o,\text{br}}$	$BW_{\text{sz}} \geq$ 9.9	$BW_{\text{nm}} \leq$ 7.8

Table 5.1: Motion parameters and the corresponding bandwidth for 3 kinds of sleep movements.

the 99-th percentile value of the speeds calculated from the dataset, which is found to be 0.33 m/s, as an estimate for the maximum possible speed of body parts during normal sleep events.[3] To estimate an upper bound for $v_{\max}$ during normal sleep events, we assume that the body part with the fastest motion is aligned with the perpendicular line to the Tx-Rx ellipse of Fig. 5.1. Hence, $v_{\max,\text{nm}} \leq 0.33$ m/s. Then, Theorem 5.2 estimates the bandwidth of the WiFi signals during a normal sleep event as $BW_{\text{nm}} = \frac{\psi v_{\max,\text{nm}}}{\lambda} + f_{o,\text{nm}}$, where $f_{o,\text{nm}}$ denotes the bandwidth of the modulating signal $v(t)$. At WiFi channel 48 and $\psi = 1$, an upper bound of this bandwidth will then be $BW_{\text{nm}} \leq \frac{0.33}{\lambda} + 2 = 7.8$ Hz.

Table 5.1 summarizes the results of our WiFi CSI bandwidth analysis during the three considered nocturnal movements: breathing, seizure, and normal sleep events. It can be seen from the table that the bandwidth of the WiFi signal during a movement can be used as a distinguishing feature that differentiates seizures from normal sleep events. We make use of this observation to design a robust nocturnal seizure detection system in the next section.

5.3 System Description

In this section, we describe our proposed framework for nocturnal seizure detection using WiFi CSI signals based on the mathematical analysis of Sec. 5.2. Fig. 5.3 shows

[3]Larger speed values are only recorded when quick jerky limb motions take place. Such events are easily identifiable and differentiable from seizures, since they typically last for less than 400 ms [110].

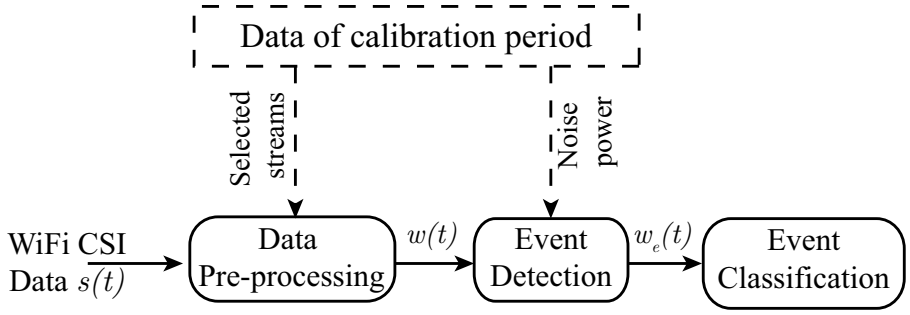


Figure 5.3: Block diagram of the proposed WiFi CSI-based nocturnal seizure detection system. The data pre-processing and event detection blocks utilize the derived WiFi CSI bandwidth during breathing (BW_{br}). The event classification module then utilizes the derived WiFi CSI bandwidth during both seizure (BW_{sz}) and normal sleep movements (BW_{nm}).

the block diagram of our proposed system, which starts by pre-processing the WiFi CSI input data to denoise the measured CSI signal and extract the part that carries the information about the human motion. Then, the denoised signal is passed to an event detection module, which decides whether the person is moving or is staying still. In case a movement event is detected (other than breathing), the CSI data during the event is then forwarded to an event classification module, which determines whether this event is a normal sleep event or a seizure. In the latter case, the system alarms the caregiver to take the necessary action. We next describe each of these components in details.

5.3.1 Data Pre-processing

As discussed in Sec. 5.1, we utilize both the CSI squared magnitude and phase difference since they both carry crucial information about the body motion. In this chapter, we consider off-the-shelf WiFi devices that can be used to extract the complex WiFi CSI information, e.g. Intel 5300 or Atheros AR9580 WiFi cards. In any of these devices, the receiver has N_R receiver antennas, which measure the WiFi CSI information on N_{sc} subcarriers. Therefore, we extract a total of $N_R \times N_{sc}$ CSI squared magnitude streams, and $(N_R - 1) \times N_{sc}$ phase difference streams (i.e. the phase difference between

each antenna and antenna 1, for all the N_{sc} subcarriers). In total, we get $N_D = (2N_R - 1) \times N_{sc}$ data streams that can be used to extract the motion information. The Intel 5300 WiFi card, for instance, has $N_R = 3$ receiver antennas and $N_{sc} = 30$ subcarriers, resulting in a total of $N_D = 150$ data streams carrying the motion information of the body. We next show how we process these N_D data streams to extract the informative part about the body motion.

Outlier Removal: We use the Hampel identifier [16] to remove the sudden and very short abrupt changes that happen in the data streams due to hardware imperfections [16].

Stream Selection: Different subcarriers on the same Rx antenna have different carrier frequencies (or wavelengths), and consequently, they undergo different levels of fading, making some subcarriers noisier than others. To enhance the system's robustness, it is then important to select only the most informative/least noisy data streams to be subsequently used in the rest of the seizure detection algorithm. In order to do so, we use the data of a short *calibration period*, in which the sleeping person is only breathing and not doing any movements or having a seizure. This one-time calibration can be easily administered by a caregiver prior to system deployment, and recalibration can be done as needed.

The stream selection algorithm works as follows. Since the calibration period is known to have only breathing motion, the CSI data contains frequency components only in the band $f \leq BW_{\text{br}}$, where BW_{br} is the maximum bandwidth for WiFi CSI during breathing, which we have shown in Sec. 5.2 to be $2f_{o,\text{br}}$. Any frequency content above BW_{br} is thus due to noise. Hence, given all the data streams in a calibration window of duration T_{cal},

we calculate the Signal-to-Noise Ratio (SNR) of the i-th data stream as follows

$$SNR_i = \frac{\sum\limits_{0<f\leq BW_{\rm br}} S_i(f)}{\sum\limits_{f>BW_{\rm br}} S_i(f)} \tag{5.10}$$

where $S_i(f) = |\sum\limits_t s_i(t)e^{-j2\pi ft}|^2$, $s_i(t)$ is the i-th data stream, $BW_{\rm br} = 2f_{o,{\rm br}}$, and $f_{o,{\rm br}}$ is the maximum normal breathing frequency, which is equal to 0.3 Hz in adults.

We then select the $N_{\rm DS}$ data streams with the highest SNRs from the calibration data, and use only this set of streams in the operation phase until after a major event happens (for instance, a seizure). The system can then re-calibrate by processing all the data streams again and re-selecting the new top $N_{\rm DS}$ streams in terms of SNR in the new person's pose/orientation. For the implementation of our system (see details in Sec. 5.4), we set $T_{\rm cal} = 13$ sec and $N_{\rm DS} = 15$.

PCA denoising: After extracting the set of the best $N_{\rm DS}$ data streams, we further denoise these streams during operation phase using Principal Component Analysis (PCA) as described in [29]. More specifically, we extract the first principal component $w(t)$ of the data, which carries the motion information since it is common to all the data streams, while the noise is distributed among all the different principal components [29].

In order to show the performance of the preprocessing module, we conduct an overnight sleep experiment, where WiFi transceivers are placed on both sides of a bed on which a subject sleeps. An accelerometer is attached to the upper right arm of the subject to collect ground truth sleep motion data. Fig. 5.4 shows a 4-hour snippet of the processed WiFi data $w(t)$ as well as the accelerometer output during the same period. A 13-second calibration period is chosen right after the subject goes to sleep and the selected streams are then used for the rest of the night. It can be clearly seen that the preprocessed WiFi data $w(t)$ carries the same motion information as the accelerometer. In the right part of

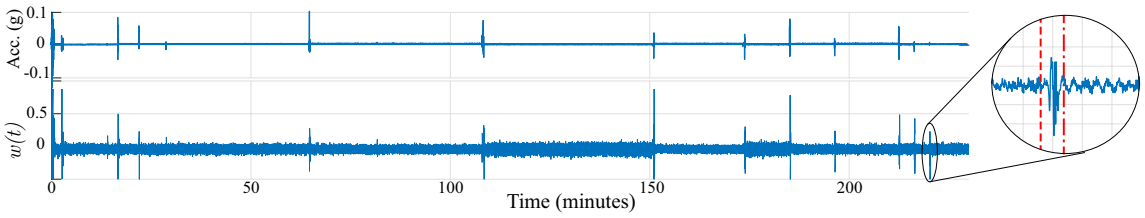


Figure 5.4: (Top) Sample output of the accelerometer attached to the arm of a subject during 4 hours of overnight sleep. (Bottom) The PCA-denoised stream $w(t)$ of the WiFi data collected during the same time period. The dashed red line in the zoomed-in part shows the start of a sample event that is detected by our event detection module, while the dashed-dotted red line indicates its end.

the figure, we zoom in to one of the movements, where the breathing signal, as well as the motion, can be clearly seen in the WiFi data.

5.3.2 Event Detection

As described in the previous section, the data pre-processing module outputs a signal $w(t)$ which is the denoised version of the CSI measurements at the receiver. This signal is then fed to an *event* detection module. By "event", we mean the state of the sleeping person engaging in any kind of *non-breathing* movement. More specifically, the movement can be normal sleep events, e.g., posture adjustments, or abnormal, e.g. a seizure. The nature of the event (whether it is normal or abnormal) will be decided in a later stage, which we shall describe in Sec. 5.3.3.

In order to detect an event in the signal $w(t)$, we use a moving window of duration $T_{\text{win}}^{\text{ED}}$. If the person was only breathing during an instance of the moving window, the signal $w(t)$ during that window will have a frequency spectrum that is concentrated below BW_{br}, as discussed in Sec. 5.3.1. On the other hand, if the person engages in any type of non-breathing movement, the signal $w(t)$ within the time window will have non-negligible frequency content above BW_{br}. Therefore, we can utilize the energy content of the spectrum of $w(t)$ above the frequency BW_{br} to indicate the presence of an

event. More specifically, let $\mathcal{H}_1$ denote the hypothesis of having an event, and $\mathcal{H}_o$ denote otherwise. To decide if there is an event at time $t = \tau$, we use the decision rule

$$\sum_{f>BW_{\text{br, adj}}} \left| \sum_t w(t)\kappa(t,\tau)e^{-j2\pi ft} \right|^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \sigma_{\text{th}}, \tag{5.11}$$

where $\kappa(t,\tau)$ is a rectangular window of length $T_{\text{win}}^{\text{ED}}$ ending at time $t = \tau$. Note that due to the time-windowing of the signal $w(t)$, the frequency spectrum of the windowed signal is that of the original signal convolved with a sinc function, which increases the bandwidth of the signal by an amount of $1/T_{\text{win}}^{\text{ED}}$. Hence, the value of BW_{br} is adjusted to be $BW_{\text{br, adj}} = 2f_{o,\text{br}} + 1/T_{\text{win}}^{\text{ED}}$, where $f_{o,\text{br}}$ is the maximum normal breathing frequency.[4]

In order to determine the value of σ_{th}, we utilize the processed data of the calibration period (whose duration is T_{cal}) described in Sec. 5.3.1 to evaluate the following,

$$\sigma_c^2 = \max_{\tau} \left\{ \sum_{f>BW_{\text{br,adj}}} \left| \sum_t w_{\text{cal}}(t)\kappa(t,\tau)e^{-j2\pi ft} \right|^2 \right\} \tag{5.12}$$

where $w_{\text{cal}}(t)$ is the processed data of the calibration period. σ_c^2 is then the maximum energy content of the calibration data above $BW_{\text{br,adj}}$, which is an estimate of the noise power in the band of $f > BW_{\text{br,adj}}$ when there is no event. We then set $\sigma_{\text{th}} = \epsilon\,\sigma_c^2$, where ϵ is a design parameter.

The zoomed-in part of Fig. 5.4 shows a sample normal sleep movement from a sleeping subject. The vertical red dashed line shows the start of the detected event using our proposed event detection module, while the vertical red dashed-dotted line shows its end. It can be seen that our event detection module was able to accurately localize the start and the end of the event.

[4]Note that for a large window size (large T_{win}), the additional bandwidth $1/T_{\text{win}}$ can be neglected with respect to the original signal bandwidth. In such cases, the bandwidth calculations need not be adjusted.

5.3.3 Event Classification

Once an event has been detected, the processed data $w(t)$ during the event is then passed to an *event classification* module that determines whether this event is normal or abnormal. As discussed in Sec. 5.2, the duration of a seizure is usually longer than that of any normal event. However, relying solely on event duration for deciding whether the event is a seizure or not induces an unfavorable delay in the system response, as the system would have to wait for a relatively long period of time before declaring an event as a seizure, which can lead to undesirable complications for the patient. It is then crucial to analyze the detected events in terms of their frequency content, using the analysis and parameters derived in Sec. 5.2, in order to have an early and robust detection. We next describe our event classification algorithm.

First, any event whose duration is less than a tolerable value $T_{\min}$ is declared as a normal event. This step is important to avoid the unnecessary computational overhead of analyzing very short events, such as sleep jerks or very quick limb movements, since it is almost impossible for a tonic-clonic seizure to have such a short duration [110]. It is noteworthy that this comes at the expense of a small delay in the response time, since a seizure would only be declared at least $T_{\min}$ after its onset. As a design choice, we set $T_{\min} = 5$ sec for our system implementation. We will show the effect of varying $T_{\min}$ on the system performance in Sec. 5.5.

For the rest of the events (whose durations are larger than $T_{\min}$), let $w_e(t)$ denote the processed CSI measurements during the event. We divide $w_e(t)$ to consecutive overlapping windows of length $T_{\text{win}}^{\text{EC}}$, and estimate the bandwidth of $w_e(t)$ as the median of the bandwidths of the signals in the overlapping windows. More specifically, the bandwidth

of $w_e(t)$ is estimated as

$$B_{w_e} = \underset{\tau}{\text{median}} \left\{ B : \frac{\sum\limits_{f>B} \left| \sum\limits_{t} w_e(t)\kappa'(t,\tau)e^{-j2\pi ft} \right|^2}{\sum\limits_{f>0} \left| \sum\limits_{t} w_e(t)\kappa'(t,\tau)e^{-j2\pi ft} \right|^2} = 0.1 \right\}, \tag{5.13}$$

where $\kappa'(t,\tau)$ is a rectangular window of length $T_{\text{win}}^{\text{EC}}$ ending at $t = \tau$, and the quantity inside the braces is the 90-th percentile bandwidth of the signal within the window ending at $t = \tau$. For an ongoing long event, the bandwidth B_{w_e} is updated by adding more time windows of the new data to the calculation of Eq. 5.13. This method of estimating the bandwidth of the signal $w_e(t)$ is favorable for real-time operation, since it requires a fixed-length FFT operation for a window of size $T_{\text{win}}^{\text{EC}}$ to update the bandwidth of an ongoing event.

We declare a seizure if the bandwidth B_{w_e} exceeds a threshold f_{th}. By using the spectral analysis of Sec. 5.2 and the corresponding bandwidth calculations of Table 5.1, we set $f_{\text{th}} = \frac{9.9+7.8}{2} = 8.85$ Hz, since this value optimally separates the bandwidths of the WiFi signal during seizures from the ones during normal events. We will study the effect of changing f_{th} on the system performance in Sec. 5.5.

5.4 Experimental Setup

In this section, we describe the experimental setup we shall use as a proof-of-concept for our proposed seizure detection system.

Experimental Setup: For the WiFi CSI data collection, we use two laptops equipped with Intel 5300 WiFi cards. One of the laptops (the Tx) transmits WiFi packets at a rate of 200 packets per second on WiFi channel 48, which has a carrier frequency of 5.24 GHz. The other laptop (the Rx) uses CSItool [79] to measure the CSI data of 30

Figure 5.5: We tested our proposed approach in 7 different locations. Four sample locations are shown here.

WiFi subcarriers on 3 Rx antennas. The CSI magnitude data and the phase difference data with respect to antenna 1 (i.e. $\angle c_2(t) - \angle c_1(t)$ and $\angle c_3(t) - \angle c_1(t)$) are then logged and processed offline using MATLAB. We collect the WiFi data in 7 different dorm rooms/bedrooms (some of which are shown in Fig. 5.5). In all the locations, we start by placing the Tx and Rx on two different sides of the bed on which the test subject lies down (with Tx-Rx distance of ~ 2.5 m). The antennas of the Tx and Rx are both elevated by 70 cm above the bed level. We then study the impact of different Tx/Rx configurations in Sec. 5.5. Note that for the Rx, external tripod-mounted antennas may be used in order to make the Rx at the same height as the Tx. This configuration for the relative positioning of the Tx, the Rx, and the bed results in $\psi \approx 1$ (the angle ϕ in Fig. 5.1 is $\sim 60^\circ$), independent of the person's pose or orientation on the bed.

Test Subjects and Experiment Protocol: We recruited a total of 20 student actors (5 females and 15 males) to participate in our experiments, where each subject participates in one or more of the experimental locations. In total, the number of subjects participating in each of the 7 locations are 11, 6, 4, 2, 1, 1, and 1 subjects, respectively.[5] Each participant was consensually trained on how to simulate a tonic-clonic seizure and shown public online YouTube videos explaining how tonic-clonic seizures look like. It is worth noting that seizure acting is a common practice in medical schools, where healthy

[5]The Institutional Review Board (IRB) committee has reviewed this research and determined that it does not constitute human subject research. Furthermore, all the experiments that were carried out during the pandemic followed the strict COVID-19 safety guidelines put in place by our institution.

persons (known as *standardized patients*) are recruited to act out different medical conditions to provide introductory training opportunities for medical students [116, 117]. Hence, testing a system on simulated seizures is an important step towards more advanced clinical trials.

For each subject, the receiver starts logging the CSI data when the subject is in a sleep state (only breathing and in any generic position) for at least 15 seconds (part of which to be used as one-time calibration data). Then the subject starts simulating seizures and normal sleep movements. Each seizure instance is simulated for at least 20 seconds.[6] In total, each participant does 10 seizure simulations per location, resulting in a total of 260 independent instances of seizure data across all the locations. Similarly, each subject performs several normal non-breathing sleep movements spontaneously in each experiment. By observing the subjects' movements, they included posture adjustments (e.g. switching from lying on their side to lying on their back), limb-only movements (e.g. stretching or tucking the knee), scratching, stretching, coughing, sneezing, and sleep jerks. Overall, we collected a total of 410 independent normal sleep events from all subjects in all the locations.

Performance Metrics: We test the performance of our system according to the following two performance metrics:

1. Seizure Detection Rate (SDR): which is defined as the number of detected seizures, divided by the total number of seizures (expressed as a percentage).

2. Probability of False Alarm (P_{FA}): which is defined as the number of normal sleep events which are incorrectly classified as seizures, divided by the total number of detected normal sleep events.

3. Response Time (RT) for seizures: which is defined as the time at which the event

[6] An actual tonic-clonic seizure can last for 1 to 3 minutes. However, it is a physically-challenging task for a healthy person to simulate it for such a long time.

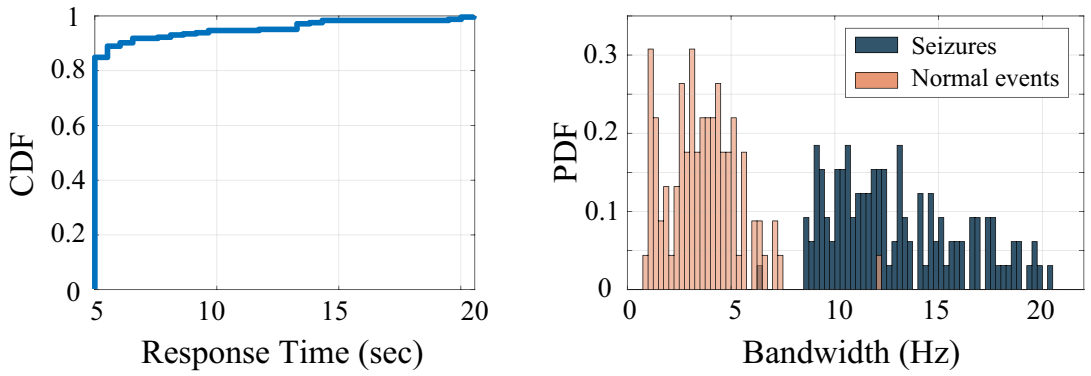


Figure 5.6: (Left) CDF of the system's response time. The mean response time is 5.69 seconds. (Right) PDF of the bandwidths of seizure events and normal sleep events, showing a gap in the 7-9 Hz band.

classification module detects the seizure, measured with respect to the seizure's onset.

Algorithm Parameter Values: We set the following values for different algorithm parameters, $T_{\text{cal}} = 13$ sec, $T_{\text{win}}^{\text{ED}} = 2$ sec, $T_{\text{win}}^{\text{EC}} = 4$ sec, $T_{\text{min}} = 5$ sec, N_{DS}=15, $\epsilon = 2$, and $f_{\text{th}} = 8.85$ Hz. The effect of varying some of these parameters on system performance is shown in Sec. 5.5.

5.5 Experimental Results

In this section, we present the performance evaluation results of our proposed seizure detection algorithm.

For the seizure data instances in all the locations, our proposed system was able to detect 244 out of the 260 seizures, resulting in a seizure detection rate (SDR) of 93.85%. It is worth noting that the event detection module was able to detect all the 260 seizures. However, the event classification module misclassified 16 out of the detected 260 events. Fig. 5.6 (left) shows the CDF of the response times of the detected seizures, showing that our system achieves a Mean Response Time (MRT) of 5.69 sec. Such an early detection is important for the caregiver to provide the needed medical assistance as soon as possible.

Paper	Modality	Seizure det. rate	MRT (sec)	$P_{\mathbf{FA}}$ /night	Non-invasive	Privacy
Avg. of [118]	Acc.	90.7%	41.1	0.3	✗	✓
[119]	Acc.	91%	17	0.1	✗	✓
[120]	Mattress	84.6%	–	–	✗	✓
[55]	Video	95%	–	1	✓	✗
[121]	Acc.+HR	96%	15	0.23	✗	✓
Our work	**WiFi**	**93.85%**	**5.69**	**0.23***	✓	✓

* Based on an average of 3 normal events per hour, for 8 hours of night sleep.
Acc = Accelerometry. HR = Heart Rate.

Table 5.2: Comparison with state-of-the-art in seizure detection.

In terms of locations, the seizure detection rate in the 7 locations was 93.6%, 95%, 90%, 90%, 100%, 100%, and 100%, while the mean response time in the 7 locations was 5.8, 5.8, 5.68, 5.58, 5.65, 5 and 5.1 sec, respectively. This shows that the system's performance is insensitive to the deployment environment, since the static multipath does not affect the information-bearing parts of the received WiFi signal, as discussed in Remark 5.2.

In terms of normal events, our event detection module was able to detect 406 out of the 410 normal events. It is worth noting that it is irrelevant if the system misses some normal events, as the main purpose of the system is seizure detection with as few false alarms as possible. Out of the detected normal events, only 4 events were incorrectly classified as seizures, resulting in a probability of false alarm $P_{\mathrm{FA}} = 0.0097$. Fig. 5.6 (right) shows the densities of the measured bandwidths of the WiFi signals during seizure events as well as normal sleep events. The distributions of the bandwidths show a clear gap in the band of 7-9 Hz, which validates the theoretical bandwidth characterization of Sec. 5.2.

Processing time: It takes 18 ms, on average, to process one second of collected data, using our algorithm of Sec. 5.3.

Comparison to state-of-the-art: [118] provides a survey for in-home tonic-clonic seizure detection algorithms that use different modalities, e.g. accelerometry, mattress units, and video, to detect tonic clonic seizures on real epilepsy patients. Table 5.2

compares the performance of our proposed system to the performance of the different detection techniques reported in the survey of [118], as well as other multimodal seizure detection papers. Overall, our results show the robustness of our proposed system, in terms of achieving a very good seizure detection rate, probability of false alarm, and a fast average response time of 5.69 seconds to detect a seizure, while being non-invasive, and privacy preserving. We note that part of the contribution of this chapter was also to develop a new mathematical model that can enable seizure detection using WiFi signals, while most of the existing work is mainly either testing an existing product, or utilizing straightforward modalities, e.g. accelerometry. Furthermore, our approach is the only privacy-preserving one that is also non-invasive. While our results are based on simulated seizures, they constitute a strong proof-of-concept for our proposed idea/mathematical models, which shows how RF signals (e.g. WiFi) can be used as a non-invasive, robust, and affordable alternative for nocturnal seizure detection. Hence, our proposed algorithms serve as a basis for a system that can be subsequently tested in clinical settings, towards the ultimate goal of making such technology available to the public.

5.5.1 Effect of varying $f_{\mathbf{th}}$ and $T_{\mathbf{min}}$

Based on our theoretical analysis of Sec. 5.2, we concluded that a threshold of $f_{\text{th}} = 8.85$ Hz optimally separates the bandwidth of the WiFi signals during normal sleep movements from that during seizures. In this section, we study the effect of varying f_{th}, while keeping all other system parameters at their default values. Fig. 5.7 (left) shows (1-SDR) and P_{FA} as a function of f_{th}. It can be seen that SDR decreases (becomes worse) when increasing f_{th}, since more seizure events can go undetected due to their bandwidth being less than the higher f_{th}. On the other hand, increasing f_{th} improves P_{FA}, since it becomes less likely for the bandwidth of the WiFi signal during a normal

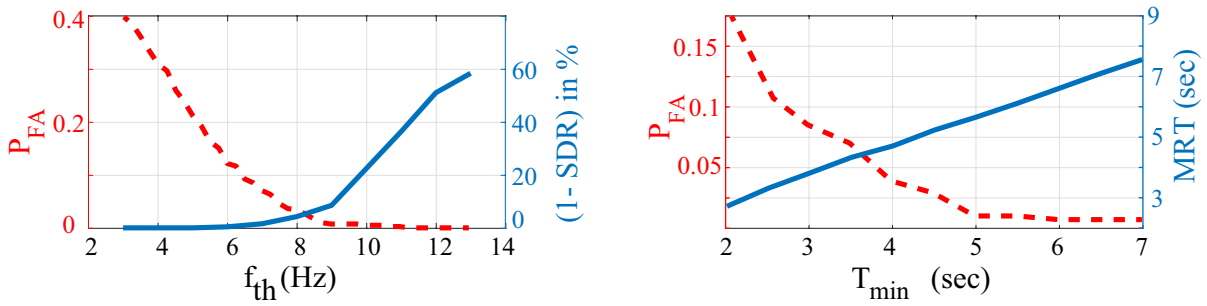


Figure 5.7: (Left) P_{FA} and (1-SDR) as a function of f_{th}: increasing f_{th} degrades SDR while improving P_{FA}. (Right) MRT and P_{FA} as a function of T_{min}: increasing T_{min} degrades MRT while improving P_{FA}.

sleep event to exceed a higher f_{th}. We can see that the mathematically-driven value of 8.85 Hz strikes a good balance between SDR and P_{FA}.

Next, we study the effect of varying T_{min}, which is the minimum duration for an event to be passed to the event classification module. Fig. 5.7 (right) shows MRT and P_{FA} as a function of T_{min}. Expectedly, increasing T_{min} increases MRT, since a higher T_{min} means that the event classification module (which determines whether the event is a seizure or not) is not activated for a longer time after the seizure onset. On the other hand, increasing T_{min} improves P_{FA}, since a higher portion of the normal events are declared normal by default due to their short duration. It can be seen that the chosen value of $T_{\text{min}} = 5$ sec strikes a good balance in the MRT-P_{FA} tradeoff. It is worth noting that SDR does not change as a function of T_{min} in Fig. 5.7, and as such is not plotted.

5.5.2 Effect of Tx-Rx positioning

For all the previous results, we considered a setting (henceforth denoted by C#1) where the position of the Tx/Rx with respect to the bed resulted in a value of $\psi \approx 1$ (see Fig. 5.5). In this section, we test our system with different Tx/Rx positions resulting in different ψs. Based on our analysis of Sec. 5.1, ψ impacts the bandwidth characterization of seizure and normal sleep movements as follows: $BW = \frac{\psi v_{\text{max}}}{\lambda} + f_o$, where v_{max} and f_o

Setting	ψ	f_{th} (Hz)	SDR (%)	MRT (sec)	P_{FA}	
C#1	1	8.85	93.85	5.69	0.0097	
C#2	1.4	11.64	90	5.67	0.008	*Different configs.*
C#3	0.7	6.69	100	6.75	0.016	
C#4	1.44	11.94	90	5	0.008	*Different heights*
C#5	1.61	13.15	100	7.3	0.016	

Table 5.3: Performance in different Tx/Rx placement settings.

denote the $v_{\max}$ and f_o of the corresponding cases.

To test the sensitivity of our system to different Tx/Rx positions and their corresponding ψ, we carry out extensive experiments on one test subject (in location 4 of Fig. 5.5) by changing either the Tx/Rx locations in the same horizontal plane (to which we refer as changing their configuration), or changing their heights.

Changing Tx/Rx configuration: In order to test the sensitivity of the system to the placement of the Tx/Rx in the horizontal plane, we conduct experiments in two additional configurations, C#2 and C#3. In C#2, the Tx and the Rx are placed on one side of the bed, such that the line connecting the Tx and Rx is parallel to the edge of the bed and 70 cm away from it. Such a configuration can be of particular interest in practical situations in which one side of the bed is not accessible, e.g. if the bed is placed next to a wall. The distance between the Tx and the Rx is 2 m, and both are elevated by 70 cm above the bed level. This setup results in $\psi \approx 1.4$, which will result in $BW_{\text{sz}} \geq$ 13.23 Hz, and $BW_{\text{nm}} \leq$ 10.06 Hz using our mathematical derivations. We thus use $f_{\text{th}} = \frac{13.23+10.06}{2} =$ 11.64 Hz for this configuration. On the other hand, in C#3, the Tx and Rx are placed on two different sides of the bed, with a Tx-Rx distance of 3.6 m, while they are elevated by 70 cm above the bed level. This setup results in $\psi \approx$ 0.7, and $f_{\text{th}} = \frac{7.36+6.03}{2} =$ 6.69 Hz.

In each of the configurations, the test subject simulates a total of 10 seizure instances

and 125 normal sleep events. We summarize the evaluation results of these experiments in Table 5.3. It can be seen that the performance of the system in C#2 and C#3 is comparable to that of the main configuration (C#1), showing that the performance of our proposed pipeline is robust to different Tx/Rx configurations.

Changing Tx/Rx heights: In order to test the sensitivity of the system to antenna heights, we conduct experiments in two additional settings, C#4 and C#5. In both settings, the Tx and Rx are placed ∼2.5 m apart on both sides of the bed (similar to C#1), but their heights are elevated to 1.3 m above the bed level in C#4, and 1.7 m above the bed level in C#5. Using simple geometry, it can be verified that in C#4, $\psi = 1.44$ ($f_{\text{th}} = 11.94$ Hz), while in C#5, $\psi = 1.61$ ($f_{\text{th}} = 13.15$ Hz). Again, in both settings, the test subject simulates a total of 10 seizure instances and 125 normal sleep events. Table 5.3 shows that the performance of the system in C#4 and C#5 is comparable to that of the other configurations, indicating that our proposed pipeline is robust to different Tx/Rx heights.

5.5.3 Multi-person Operation

In-home seizure detection systems are primarily designed for caregivers who do not share the same bed (or bedroom) as the patient, since, otherwise, they would be alerted by the patient's seizure movements. However, in order to show the robustness of our proposed system, we next show that it can still be deployed in a multi-person setting where multiple people share the same bed. In such a case, the event detection module detects any movement done by any of the sleeping persons. In order to test this, we conducted a 10-minute experiment where two people lie down next to each other on a bed. Person 1 simulates seizures at the 6 and 9-minute marks. Otherwise, both people frequently simulate normal sleep movements. As such, there are a number of instances

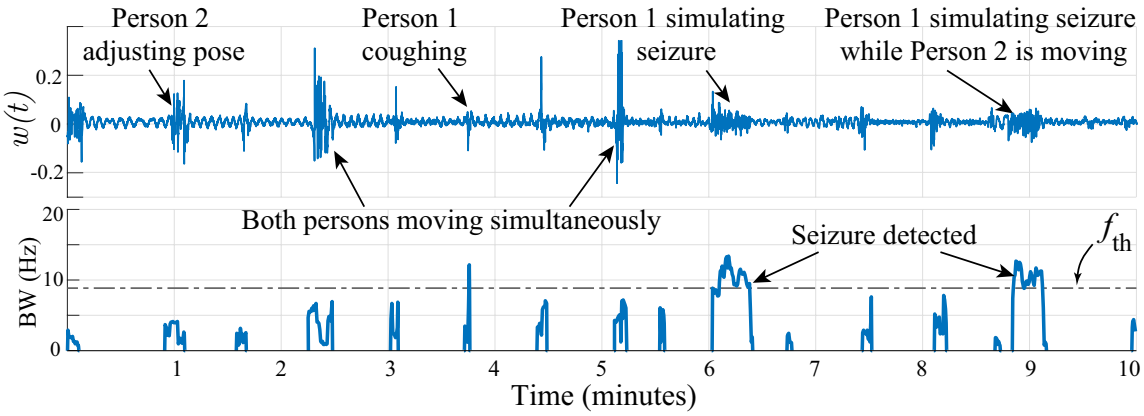


Figure 5.8: (Top) The PCA-denoised data $w(t)$ in a 10-minute experiment with 2 subjects. (Bottom) The bandwidth of $w(t)$ during the detected events. It can be seen that the seizures are the only events whose bandwidth exceeds f_{th} for an extended period of time.

where both people move at the same time, or one person moves normally while the other one is simulating a seizure. Fig. 5.8 (top) shows the PCA-denoised stream $w(t)$, in which perturbations are clearly visible whenever either persons engages in any kind of movement, while Fig. 5.8 (bottom) shows the bandwidth of the WiFi signals during the detected events. It can be seen that the bandwidth exceeds f_{th} for an extended period of time only during the seizure instances, which are correctly classified as seizures, even though the second person was moving during the second seizure instance. Otherwise, during normal movements, even if both persons are moving simultaneously, the events are not classified as seizures.

5.6 Discussions

Robustness to movements by other people: In Sec. 5.5.3, we have shown that our proposed system is robust to movements by other sleeping people in the same environment, since their normal sleeping movements have the same characteristics as those of the patient. Next, consider the case where other simultaneous movements happen, such as

those of a walking person. The spectrogram of the signal reflected off of a walking person has specific characteristics. Thus, as part of future work, one can study the differentiability of the signals induced by walking from those induced by seizures. Furthermore, the reflected signals off of other moving targets can also be filtered out at the Rx by exploiting more signal dimensions. For instance, multiple antennas at the Rx can separate the received signals based on their Angle-of-Arrival (AoA).

Real-time processing: While the results presented in this chapter are based on the offline processing of the collected CSI data, the design of the algorithm has taken real-time processing considerations into account. For instance, the steps of the algorithm, such as event detection and bandwidth estimation, rely on moving windows, instead of processing all the data at once, which facilitates the implementation of the algorithm in real-time. Moreover, our proposed algorithm is computationally efficient, taking only 18 ms on average to process one second of the data. Since the window size of the event classification module is 4 sec, this means that the additional processing delay to the system's response time (whose mean is 5.69 sec) is only $18\times4=72$ ms.

Clinical trials: In this chapter, we proposed the first RF-based system for nocturnal seizure detection, by developing mathematical models that can enable this. We also validated our proposed approach by extensive experiments on seizures simulated by actors. The results of this preliminary validation show the great potential of using WiFi signals as an appealing alternative to the currently available products, which are costly, uncomfortable, or unreliable. Towards the ultimate goal of making this technology available to the public, the next step is to develop a prototype of the proposed system, which can then undergo extensive clinical trials on real patients, and become available to the epilepsy patients and their caregivers.

Chapter 6

Counting a Stationary Crowd Using Off-the-Shelf WiFi

In this chapter, we present a new foundation for counting a stationary (seated) crowd using a pair of WiFi transceivers. A case where a number of seated people are engaged in an activity is applicable to many real-world settings, such as the audience of different kinds of social events (e.g. seminars, presentations, lectures), the crowd in a movie theater, or the crowd in a wedding ceremony. It also applies to situations in which each person is separately engaged in an individual task, such as readers in a library. Fig. 6.1 shows a few real-world examples of a stationary crowd.

Here is our underlying proposed idea. Consider a scenario such as the ones shown in Fig. 6.1, where a number of people are seated. The people in the crowd are stationary, i.e. with no major body motion except breathing. However, people do not stay still for a long period of time and frequently engage in different kinds of small in-place natural body motions (called fidgets [122]), such as adjusting their seating position, crossing their legs, checking their phones, scratching, stretching, and coughing, among many others. We then propose that the aggregate natural body fidgets of the crowd carry crucial information on

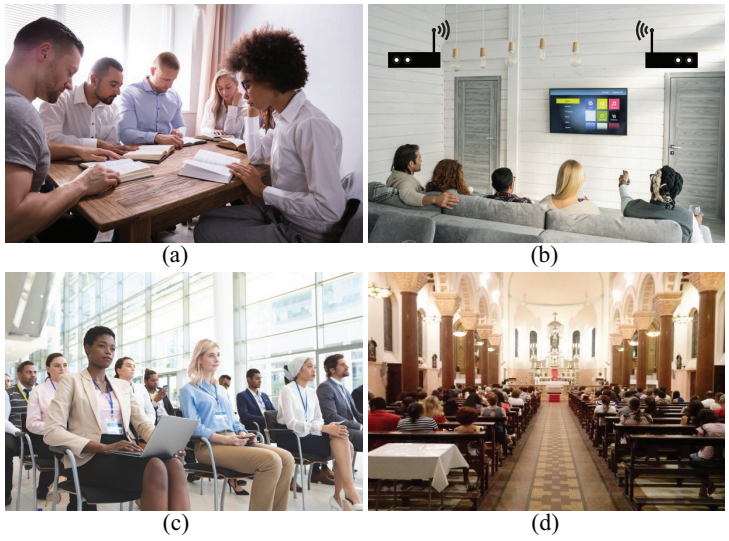


Figure 6.1: Sample application scenarios of our stationary crowd counting system: (a) a group of people reading, (b) a group of people watching a movie, (c) attendance of a presentation, and (d) attendance of a religious event.

the crowd count. Furthermore, we develop a new mathematical characterization for the statistics of the *Crowd Fidgeting Periods (CFPs)*, which we define as the periods in which at least one person in the area is fidgeting, as well as *Crowd Silent Periods (CSPs)*, which we define as the periods in which no one in the area is fidgeting (i.e. everyone is only breathing), and show their dependency on the total number of people in the area. We then demonstrate that a Maximum A Posteriori (MAP) estimator of the number of people can be obtained using these new mathematical models. Our mathematical characterizations are inspired by a 1985 queuing theory paper [123]. More specifically, we show how our problem of mathematically modeling the statistics of the crowd fidgeting and silent periods can be posed similar to an $M/G/\infty$ queuing theory problem. By borrowing and adapting mathematical tools from queuing theory, we can then characterize the aggregate crowd fidgeting dynamics and mathematically relate them to the total number of people.

6.1 A Mathematical Model for Crowd Fidgeting

In this section, we develop a new mathematical model to describe the statistics of the fidgets of a group of stationary people, and show how they relate to the total number of people. Consider a scenario such as the ones shown in Fig. 6.1, where N people are seated in a WiFi-covered area. The people in the crowd are stationary, i.e. staying still with no major body motion except breathing. However, as discussed earlier, people do not stay still for a long period and frequently engage in different kinds of small in-place motions, called fidgets.

Let $t = 0$ denote the start of the measurement time. Let $t_i^{\{n\}}$ denote the start of the i-th fidget of the n-th person, for $n = 1, 2, \ldots, N$. The n-th person then fidgets for a duration of $d_i^{\{n\}}$ before returning to the state of being stationary, i.e. $d_i^{\{n\}}$ is the time duration of the i-th fidget of the n-th person. $d_i^{\{n\}}$ can be modeled as an independent and identically distributed (i.i.d) random variable, with a Probability Distribution Function (PDF) $p_D(d)$. Let $T_i^{\{n\}}$ be the time between the start of the i-th fidget of the n-th person and the start of his/her $(i+1)$-th fidget (i.e. $T_i^{\{n\}} = t_{i+1}^{\{n\}} - t_i^{\{n\}}$). Let $p_n(T)$ denote its distribution. Fig. 6.2 presents a visual demonstration of the aforementioned quantities.

Similar to many processes in different scientific fields, the natural fidgeting process of an individual is well modeled by a Poisson process [124].[1] Therefore, the inter-fidget times of the n-th person, $T_i^{\{n\}}$, will have an exponential distribution, i.e. $p_n(T_i^{\{n\}} = T) = \frac{1}{\gamma_n} e^{-T/\gamma_n}$, where γ_n is the average inter-fidget time of the n-th person. Note that we do not assume or require that all people fidget at exactly the same rate, i.e. γ_n's can be different. Instead, we take γ_n to be a random variable taken from a distribution $p_\Gamma(\gamma)$. We shall discuss $p_\Gamma(\gamma)$ and $p_D(d)$ in more detail in Sec. 6.3.

It is well-known that the superposition of N different Poisson processes results in

[1] The validity of the assumption that the individual fidgeting process can be modeled as a Poisson process will be discussed in more detail in Sec. 6.5.

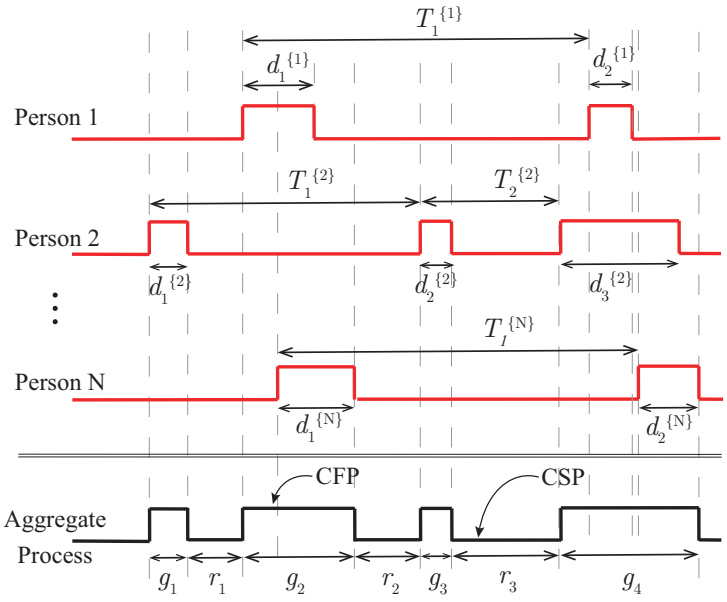


Figure 6.2: A sample fidgeting timeline of N people in an area, where a non-zero signal indicates a fidget. $d_i^{\{n\}}$ denotes the duration of the i-th fidget of the n-th person, while $T_i^{\{n\}}$ represents the time between the i-th and $(i+1)$-th fidgets of the n-th person. An aggregate fidgeting process results from the superposition of the individual fidgeting processes, where the state of no one fidgeting is referred to as the *Crowd Silent Period (CSP)* (with a duration r), while the state of at least one person fidgeting is referred to as the *Crowd Fidgeting Period (CFP)* (with a duration g). A sample CFP and CSP are marked on the figure.

a Poisson process whose rate is the sum of the rates of the individual processes [125]. Therefore, to analyze the collective fidgeting behavior of the crowd, denote by $T^{[c]}$ the inter-fidget time of the overall crowd fidgeting process (i.e., the time between the start of any two consecutive fidgets happening from any individuals in the crowd). $T^{[c]}$ then follows the exponential distribution:

$$p_{T^{[c]}}(T) = \left(\sum_{n=1}^{N} \frac{1}{\gamma_n} \right) e^{-T\left(\sum_{n=1}^{N} \frac{1}{\gamma_n}\right)}. \tag{6.1}$$

It should be noted that, since multiple people can fidget at the same time and each fidget has a non-zero duration, we cannot directly measure $T^{[c]}$ from the aggregate crowd fidgeting process. We next formally introduce two key parameters related to the overall aggregate fidgeting process of the crowd, which are measurable:

1. **Crowd Fidgeting Period (CFP):** we define a CFP as the state of having *one or more* individuals in the crowd engaged in a fidgeting movement. The duration of the i-th such period is denoted by g_i, and has a distribution $p_G(g)$.

2. **Crowd Silent Period (CSP):** we define a CSP as a continuous stretch of time where *no person* is engaged in any kind of fidgeting movement, i.e. all people in the crowd are only breathing. The duration of the i-th such period is denoted by r_i, and has a distribution $p_R(r)$.

Examples of CFPs and CSPs are shown in the aggregate process of Fig. 6.2. Intuitively, the PDFs $p_G(g)$ and $p_R(r)$ depend on the number of people N. For instance, when N increases, CFPs tend to get longer, due to the higher probability of overlap between fidgets of individual people. Moreover, increasing N also results in a shorter CSP. Next, we show how to mathematically characterize $p_G(g)$ and $p_R(r)$ as a function of N.

M/G/∞ Queues

Queuing theory is a branch of mathematics that studies waiting lines in systems that involve the arrival of entities (referred to as customers), which require a service from an entity that includes a number of servers [126]. Considerable research has been conducted in this area to characterize several quantities related to this problem, such as the statistics of the waiting time of the customers, the average number of customers waiting at any time instant, and other related parameters. A problem in queuing theory is traditionally denoted by a triad of symbols A/B/C, where symbol A describes the arrival process of customers, symbol B describes the distribution of the service time of a customer (the time a customer spends being served by one of the servers before leaving), and C describes the number of servers. A queuing model that is of particular interest to us (as we shall explain shortly) is an $M/G/\infty$ queue. In this queuing model, the arrival process is assumed **M**arkovian, e.g. can be modeled as a Poisson process, resulting in the inter-arrival times following an exponential distribution, the service times can follow any **G**eneric distribution, and the number of servers is ∞. An immediate characteristic of such a queue is that there are no waiting lines, since any arrival can be instantaneously served by one of the infinite servers. In 1985, ref. [123] studied two important quantities relevant to an $M/G/\infty$ queue: the busy and idle periods. A busy period is defined as a period in which there is at least one customer being served in the system, while an idle period is a period in which no customer is being served (i.e., all servers are idle).

While assuming an infinite number of servers may not be of much practical interest to most queuing problems, it is of great interest to our fidget-based crowd counting problem. More specifically, a careful inspection of the structure of the $M/G/\infty$ queue shows a strong analogy to our stationary crowd counting model. Consider the arrival process corresponding to the aggregate process of N people. As discussed earlier, this is a

Poisson process, with the inter-event times characterized by an exponential distribution as shown in Eq. 6.1. A CSP in our crowd counting model is then analogous to an idle period in the $M/G/\infty$ queue. A person starting a fidget is analogous to a customer arriving at a server station, while the fidget duration is analogous to the service time of the person. Note that the overlap of the fidgets from people is similar to the case where multiple customers are being served at the same time. Since there is no delay in fidgeting, the queue model with infinite servers then well models our case.

As such, we can tap into the literature on $M/G/\infty$ queues to mathematically characterize our CFPs and CSPs. More specifically, it can be shown that given a crowd of N people, due to the memoryless property of the exponential distribution, the duration of a CSP follows an exponential distribution whose rate is the sum of the rates of the individual people [123]. In other words, $p_R(r)$ can be written as follows:

$$p_R(r|N,\bar{\gamma}) = \frac{N}{\bar{\gamma}} e^{-\frac{rN}{\bar{\gamma}}}, \tag{6.2}$$

where $\bar{\gamma}^{-1} = \frac{1}{N}\sum_{n=1}^{N}\gamma_n^{-1}$ is the average of the fidgeting rates of people.

Additionally, it can also be shown that, given a crowd of N people, the PDF of the duration of a CFP, $p_G(g)$, is characterized as follows [123]:

$$p_G(g|N,\bar{\gamma}) = \frac{d}{dg}\left[-N\bar{\gamma}^{-1}\sum_{i=1}^{\infty} m^{*i}(g|N)\right], \tag{6.3}$$

where m^{*i} is the i-fold convolution of m with itself, and

$$m(g|N) = \frac{d}{dg}\left[-\exp\left\{-\frac{\bar{\gamma}}{N}\int_0^g (1-P_D(x))dx\right\}\right], \tag{6.4}$$

where $P_D(x) = \int_0^x p_D(u)du$ is the Cumulative Distribution Function (CDF) of an individual's fidget duration.

MAP estimation of N

The previous analysis shows how the distributions of the crowd fidgeting period, $p_G(g)$, and the crowd silent period, $p_R(r)$, are dependent on the number of people N. We then utilize this mathematical characterization and propose a Maximum A Posteriori (MAP) estimation rule to estimate the number of people N at time t, given the durations of all the CFPs and CSPs prior to t. We will show how these fidgeting/silent periods can be extracted from the ambient WiFi signals in Sec. 6.2. More specifically, let $g_1, g_2, \ldots, g_{N_f(t)}$ be the durations of all the CFPs before time t, where $N_f(t)$ is the total number of these fidgeting periods up to time t. Similarly, let $r_1, r_2, \ldots, r_{N_r(t)}$ be the durations of CSPs before time t, where $N_r(t)$ is the total number of such silent periods. The MAP estimation rule for the number of people can be written as [127],

$$\begin{aligned}
\hat{N}(t), \hat{\bar{\gamma}} &= \arg\max_{N,\bar{\gamma}} \; p(N, \bar{\gamma} | g_1, ..., g_{N_f(t)}, r_1, ..., r_{N_r(t)}) \\
&= \arg\max_{N,\bar{\gamma}} \; p(g_1, ..., g_{N_f(t)}, r_1, ..., r_{N_r(t)} | N, \bar{\gamma}) p(N) p(\bar{\gamma}) \\
&= \arg\max_{N,\bar{\gamma}} \; p_\Gamma(\bar{\gamma}) \prod_{i=1}^{N_f(t)} p_G(g_i | N, \bar{\gamma}) \prod_{j=1}^{N_r(t)} p_R(r_j | N, \bar{\gamma}),
\end{aligned} \tag{6.5}$$

where $p(.)$ denotes the probability of the argument. The last step follows from the independence of the durations of the CFPs and CSPs, and the fact that we do not assume any prior knowledge on the number of people N, i.e. $p(N)$ is taken as uniform. We also use the general PDF of p_Γ for the prior distribution on $\bar{\gamma}$. In Sec. 6.3, we show how to obtain the priors p_Γ and p_D. As we shall see, we do not need to make any prior WiFi measurements to estimate these. In summary, Eq. 6.5 allows us to estimate the total number of stationary people based on their natural fidgets.

6.2 WiFi Processing Pipeline

In Sec. 6.1, we have shown that one can estimate the number of stationary people in an area, given the durations of the CFPs $(g_1, g_2, \ldots, g_{N_f})$ as well as those of CSPs $(r_1, r_2, \ldots, r_{N_r})$. In this section, we show how to extract these periods from the ambient WiFi signal in the area of interest.

Consider the scenarios shown in Fig. 6.1, where a WiFi transmitter (Tx) transmits WiFi signals that are reflected off of the bodies of the N people in the area, after which they are received by a WiFi receiver (Rx). Let $c(t)$ denote the complex baseband received signal at the Rx, as a function of time. It has been shown, in the RF sensing literature, that the frequency content/bandwidth of the received WiFi signal increases when the speed of the moving person/object increases [128, 1, 129, 130, 30]. We make use of this fact to extract the CFPs and CSPs from the WiFi signals as follows.

When a stationary person is not fidgeting, the only body movement is the slow sinusoidal breathing motion of the chest and abdomen. Since the speed of the body motion during fidgeting is typically considerably higher than the speed during breathing, we expect the frequency content of the measured WiFi signal during fidgeting to be considerably higher than that of normal breathing, with a high probability. More specifically, it has been shown in the literature that the maximum chest displacement of a person during respiration is about 5 mm [48]. Considering that the maximum normal breathing rate of adults is $f_{\text{br}} = 0.3$ Hz [131], this chest displacement translates to a maximum instantaneous chest speed of 0.01 m/s. On the other hand, when a person is engaged in any kind of non-breathing in-place motion (e.g. fidgeting), the instantaneous speed of the body parts can increase significantly. For instance, the authors of [132] attached accelerometers to the wrists of 20 subjects to analyze their motion while doing various tasks, one of which was sitting down and relaxing, and published the acceleration data in

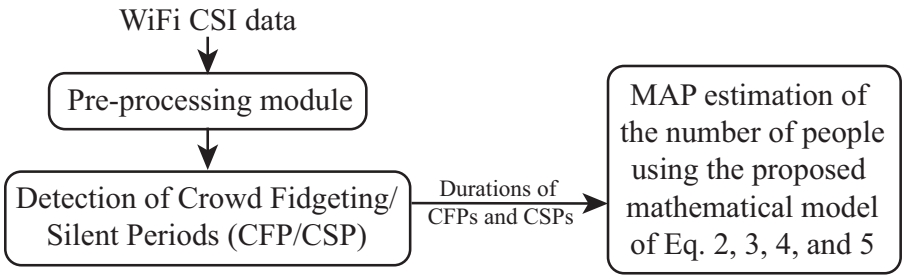


Figure 6.3: High-level overview of the proposed pipeline.

the PhysioNet database [133]. We studied this dataset and calculated the speed of body motion during the relaxing periods from their published acceleration data. We found that, during fidgets, the speed of the body motion is larger than 0.01 m/s (which is the maximum body speed during breathing) 90% of the time. Furthermore, the speed of the body motion during fidgets can take much larger values, e.g. 5 times the maximum body speed during breathing (or larger) for 80% of the time. It can even reach speeds as high as 3 m/s (i.e. 300 times larger). As such, with a high probability, the overall received WiFi magnitude or phase difference signal during a fidget will have a considerably higher frequency content than the one when only breathing. Thus, we can easily extract the fidget-related content by properly high-pass filtering the received WiFi signal (phase difference or magnitude) above B_{br}, where B_{br} is the maximum bandwidth of the received WiFi signal when only breathing. See Chapter 5 for an exact characterization of the bandwidth during breathing as well as during normal body movements. We shall use the corresponding derivations of Chapter 5 in Sec. 6.3 when characterizing B_{br}. As we shall see, B_{br} is relatively small at WiFi frequencies since the breathing rate of adults is around 0.3 Hz.

Fig. 6.3 shows our end-to-end proposed pipeline, including the pre-processing, fidget detection, and estimation steps. More specifically, we first feed the WiFi CSI phase difference between the Rx antennas to a pre-processing module. We utilize only the CSI

phase difference data since it has been shown to be more stable and robust in different deployment environments than the CSI magnitude data [49].[2] The pre-processing module extracts only a subset of the WiFi subcarriers that are less noisy, and further denoises these subcarriers by means of Principal Component Analysis (PCA), as we shall describe in Sec. 6.2.1. The denoised data is then fed to the CFP detection module, which decides whether there is any fidgeting or not based on the spectral content of the denoised data, as we shall describe in Sec. 6.2.2. The resulting fidgeting sequence is then used for counting, based on the proposed mathematical framework of Sec. 6.1. We next describe these steps in more detail.

6.2.1 Pre-processing Module

The first step in our proposed WiFi processing framework is to extract the phase difference data between the Rx antennas. More specifically, off-the-shelf WiFi Network Interface Cards (NICs), such as the Intel 5300 NIC, provide the CSI measurements on three Rx antennas, for a set of $N_{\text{sub}} = 30$ subcarriers. The absolute phase measurements on each of the Rx antennas are usually corrupted by multiple sources of errors, such as the carrier frequency offsets and sampling time offsets [17], rendering the absolute phase measurements unreliable. However, since the Rx antennas share the same oscillator, the phase difference between the Rx antennas is stable, experiences little noise, and is robust to different deployment environments [49]. As such, we extract the phase difference data between the Rx antennas $z_{2,1,k}(t)$ and $z_{3,1,k}(t)$, where $z_{i,j,k}(t) = \angle c_i(t;k) - \angle c_j(t;k)$ is the phase difference between antenna i and antenna j on the k-th subcarrier, resulting in a total of $2N_{\text{sub}} = 60$ streams of data to be utilized for fidget detection. We then remove the DC component from each data stream since the DC term only has the impact of the

[2]We note, however, that the CSI magnitude data follows the same model and can similarly be used, if stable enough, for fidget detection.

static objects in the environment (as shown in Chapters 2 and 3), and is not relevant to our analysis, which relies on the impact of the fidgets.

Stream Selection: Different subcarriers on the same Rx have different carrier frequencies (or wavelengths), and consequently, experience different levels of noise depending on the environment, making some sub-carriers noisier than others. It is then important to select only the least noisy data streams and subsequently use them in the rest of the processing pipeline. In order to do so, we utilize the measured phase difference data in a short (e.g., less than 30 seconds) one-time calibration phase prior to the real experiments. In this prior calibration phase, WiFi measurements are collected while a number of people are seated in the area and are only breathing, without any fidgeting, for a short period of time. Note that the number of people in the calibration phase is decoupled from the number of people during the real experiments and, as such, can be as low as needed. Furthermore, people can sit in any configuration. If the number of people in the calibration phase is small, e.g., 1 or 2, and the area is large, we find it better if they sit in a couple of different random configurations to pull the data of them together and find the best streams. For instance, 1 person can sit in two different locations and 2 people can sit in random locations, amounting to three 10-second data collection periods in which the subjects are only breathing. Consequently, the spectral content of the data streams in this calibration period should be confined to the band $[0, B_{\text{br}}]$ Hz. As such, we can calculate the Signal-to-Noise Ratio (SNR) for all the $2N_{\text{sub}}$ data streams as the ratio of the spectral energy content below B_{br} to the spectral energy content above B_{br}. More specifically, the SNR of the phase difference between the i-th and the j-th Rx antennas at the k-th subcarrier is computed as follows, in the calibration phase when there is only

breathing:

$$SNR_{i,j,k} = \frac{\int_{0}^{B_{\text{br}}} \left| \int z_{i,j,k}(t) e^{-j2\pi ft} dt \right|^2 df}{\int_{B_{\text{br}}}^{\infty} \left| \int z_{i,j,k}(t) e^{-j2\pi ft} dt \right|^2 df} \tag{6.6}$$

where $z_{i,j,k}(t)$ is the phase difference between the i-th and the j-th Rx antennas at the k-th subcarrier, as a function of time, and B_{br} is the maximum bandwidth of the WiFi signal during breathing. Note that, in the calibration phase (i.e. only breathing), the numerator of Eq. 6.6 contains the reflected signals, while the denominator contains only the measurement noise. We then select the top 10 data streams in terms of their SNR and use only these 10 data streams in the rest of the operation phase.

PCA denoising: After extracting the top 10 data streams (in terms of the SNR), we extract the first principal component, $w(t)$, of these data streams using Principal Component Analysis [29]. It has been shown in the literature that the first principal component contains the motion information of the moving subjects, while noise is distributed among different principal components [29]. As such, $w(t)$ serves as the denoised WiFi data that contains the motion information of the crowd.

6.2.2 Crowd Fidgeting Period (CFP) Detection

As described earlier, during CSPs, all people in the WiFi area are only breathing. Hence, the spectral content of the denoised WiFi data, $w(t)$, in CSPs is concentrated below B_{br}. On the other hand, during CFPs, at least one person in the WiFi area is fidgeting, resulting in the spectral content of $w(t)$ to span a wider band, e.g. $w(t)$ would have a bandwidth higher than B_{br}. We utilize this observation to detect the CFPs as follows. We first filter the WiFi data $w(t)$ by passing it through a high-pass filter with a cut-off frequency of B_{br}. The filtered signal $w^{\text{filt}}(t)$ would contain only noise during CSPs, while during CFPs, it contains motion data as well. Let $\mathcal{H}_1$ denote the hypothesis

that there is at least one person fidgeting in the WiFi area, while $\mathcal{H}_o$ denotes otherwise. Accordingly, we detect whether there is any fidgeting at time τ by thresholding the moving variance of $w^{\text{filt}}(t)$ as follows,

$$\text{VAR}_\tau \left\{ w^{\text{filt}}(t) \right\} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \sigma_{\text{th}}, \tag{6.7}$$

where $\text{VAR}_\tau\{.\}$ is the variance of the signal in a window of length T_{win} ending at time $t = \tau$, and σ_{th} is a threshold representing the noise floor. In order to determine the value of σ_{th}, we use the denoised WiFi signal during the calibration period, to which we refer as $w_{\text{cal}}(t)$. Since the spectral content of $w_{\text{cal}}(t)$ is concentrated below B_{br}, the filtered signal, $w_{\text{cal}}^{\text{filt}}(t)$, contains only noise. As such, we estimate the noise floor σ_{th} as follows,

$$\sigma_{\text{th}} = \max_{\tau} \text{VAR}_\tau \left\{ w_{\text{cal}}^{\text{filt}}(t) \right\}. \tag{6.8}$$

In order to show the performance of our proposed WiFi pre-processing and fidget detection modules, we conduct a 3-minute experiment where two people sit together to watch a movie. In order to get ground-truth fidgeting data, we attach a smartphone to the upper right arm of each of the two people and log the accelerometer data of the smartphones. Fig. 6.4 shows the logged accelerometer data of the two people, as well as the high-pass filtered WiFi signal $w^{\text{filt}}(t)$ extracted during the 5-minute experiment. The data in a prior 10-second calibration phase is used to extract the set of best data streams, and to estimate the value of the noise floor. It can be seen that $w^{\text{filt}}(t)$ experiences high-amplitude variations whenever any person engages in a fidgeting movement, as confirmed by the accelerometer data. The figure also plots the binary output of our CFP detection module, where the CFPs are accurately captured based on the moving variance of $w^{\text{filt}}(t)$.

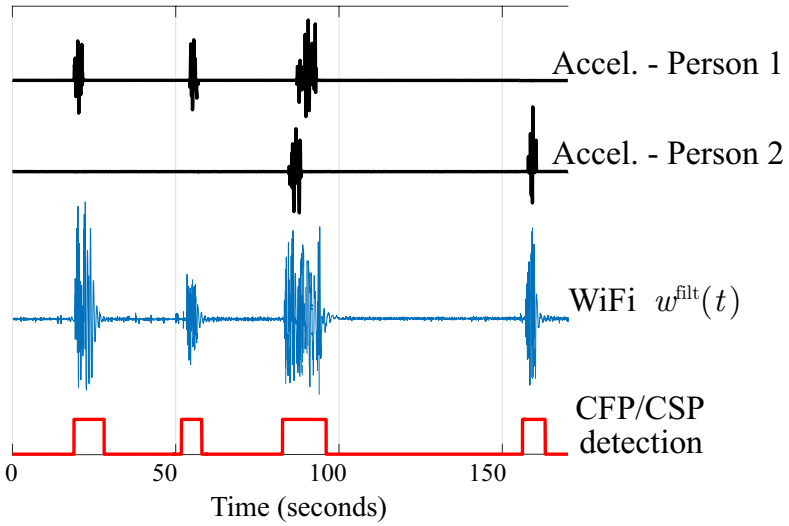


Figure 6.4: An example of CFP/CSP detection: Two people sit together to watch a movie while accelerometers are attached to their arms. The figure shows (from top to bottom) the accelerometer data of person 1, the accelerometer data of person 2, the filtered WiFi signal $w^{\text{filt}}(t)$, and the binary output of the CFP detection module. Note that people can have overlapping fidgets, an example of which is also shown in the figure. Our proposed pipeline can accurately detect and localize the CFPs/CSPs, including the ones which contain overlapping fidgets from multiple people.

6.2.3 Estimation of the number of people

Once we have detected the CFPs and CSPs, we use the proposed mathematical framework of Sec. 6.1 to estimate the total number of people, as shown in Fig. 6.3. More specifically, to estimate the number of people at time t, we extract all the durations of the CFPs (hypothesis $\mathcal{H}_1$) as well as the durations of the CSPs (hypothesis $\mathcal{H}_o$) that occurred prior to time t. Let $g_1, g_2, \ldots, g_{N_f(t)}$ denote the durations of all the detected CFPs before time t, while $r_1, r_2, \ldots, r_{N_r(t)}$ denote the durations of the CSPs before time t. Then, given the prior PDFs of a general individual's fidget duration, p_D, and average inter-fidget time, p_Γ, we can estimate the number of people using Eq. 6.2–6.5. In Sec. 6.3, we show how to obtain the prior PDFs of an individual, p_Γ and p_D. As time progresses, new outputs of the CFP detection module can be used to update the list of CFP and CSP durations, and consequently update the estimate of the number of people.

Let $\hat{N}(t)$ denote the estimate of the number of people at time t. As we collect more measurements, $\hat{N}(t)$ will start to converge to the true value. However, it may still oscillate a bit around the true value. As such, we employ a moving average filter to smooth out $\hat{N}(t)$. In other words, our final estimate for the number of people at time t is the mean of $\hat{N}(t)$ in a window of length T_{avg} seconds prior to time t.

6.3 Experimental Setup

In this section, we describe the experimental setup we use to validate our proposed framework for counting a stationary crowd.

Experiment Details: For the WiFi Tx and Rx, we use two laptops equipped with Intel 5300 WiFi NICs. The Tx laptop broadcasts WiFi packets on channel 36 (whose center frequency is 5.18 GHz), with a rate of 50 packets per second, using one Tx antenna. The Rx laptop monitors the channel and logs the CSI data of the received packets on the 3 antennas of the Rx using CSItool [79]. The logged CSI data is then processed according to the proposed pipeline of Sec. 6.2 using MATLAB. More specifically, we extract the phase difference data with respect to antenna 1 of the Rx (i.e. $z_{2,1}$ and $z_{3,1}$). The phase difference data is then pre-processed and the CFPs/CSPs are extracted using the proposed fidget detection module of Sec. 6.2.2. Furthermore, we have $B_{\text{br}} = 2f_{\text{br}} = 0.6$ Hz for a carrier frequency of 5.18 GHz, where f_{br} is the breathing rate (see Chapter 5 for more detailed derivations), $T_{\text{win}} = 2$ sec, and $T_{\text{avg}} = 90$ sec. Finally, the number of people is counted using the proposed mathematical model of Sec. 6.1, and is updated once every second.

Experiment Protocol: We carry out counting experiments in four different environments, including both through-wall and non-through-wall settings. In each of the test environments, several experiments with different number of people, different seating con-

figurations, and different activities are conducted. Each experiment lasts for 5 minutes (the estimator of the number of people typically converges much faster, as we shall see). In each experiment, a set of N subjects are asked to sit in the test area in rows of chairs, and to engage in some activity, such as watching a movie, attending a lecture, or reading. For instance, Fig. 6.7 (a) shows two sample shots of experiments where 9 subjects sit in 3 rows of chairs watching a movie on TV, while Fig. 6.8 (a) shows an experiment where 4 people sitting in two rows of chairs are watching an online lecture. People were told to act casually and normally during the experiment and just engage in the activity as they normally would. In each experimental area, a prior 10-30 second data is collected for calibration purposes (see Sec. 6.2.1 for more details on this). In total, we have conducted 47 experiments across the four environments.

Prior PDFs for p_Γ and p_D: The proposed crowd counting mathematical method introduced in Sec. 6.1 utilizes the PDF of an individual's average inter-fidget time (p_Γ) as well as the PDF of an individual's fidget duration (p_D), in order to calculate the distributions of the CFPs (p_G), and the distribution of the CSPs (p_S) of the aggregate process, as described by Eq. 6.2 and Eq. 6.3. These prior distributions describe the behavior of a general individual and are not specific to a certain group of people. They are also independent of the number of people in the crowd. We can easily get such prior fidget data of a single person from online available videos of relevant stationary activities. In this manner, we do not need to collect any prior WiFi measurements to estimate these priors. In general, similar events (e.g., attending a lecture and listening to an officiant at a wedding ceremony) will have similar priors and Youtube videos of them can be lumped together, or used interchangeably, to generate the prior PDFs. However, if events are fairly different, in terms of the required attention span, it will be more accurate to acquire different priors for them, as opposed to lumping all such videos

together. For instance, people tend to fidget less frequently when engaged in activities that require more attention, such as reading, as compared to watching a movie. Given that the general high-level type of an event would be known for a given area and given the abundance of online videos pertaining to many different activities, acquiring these prior PDFs from online videos would thus be straightforward.

In order to cover a variety of activities, in this chapter we then consider two main broad categories of activities: audience-style activities and reading-related activities. We then find the prior PDFs of an individual's fidget duration (p_D) and average inter-fidget time (p_Γ) by utilizing relevant online videos for both activities and logging the time-stamps of fidget start and end times of an individual in the video. We next elaborate in more details.

• Crowd as an audience: this category covers a wide spectrum of real-world social gatherings, where the attention of a group of people is focused on one main source of interest. Examples of this category include the audience of a lecture, presentation, or a seminar, the audience in a cinema/theater or a home-movie setting, or the audience of a wedding ceremony. For this category, we have collected 14 public online Youtube video of different lectures/presentations, with an average video length of 30 minutes, and extracted individual fidgets of a total of 30 individuals.[3]

• Crowd of readers: this category covers any setting (such as a library), where each individual in a group of people is reading. For this category, we have collected 24 public online Youtube videos of people reading (under the search keyword "Read With Me"),[4] and extracted fidget data of a total of 24 individuals. Note that in addition to the examples of fidgets we already mentioned (such as pose adjustments, checking cell phones,

[3]Sample videos we have used for this category can be found at: `youtu.be/r_w7pfulsn8` and `youtu.be/nfrmH65kS-E`.

[4]Sample videos we have used for this category can be found at: `youtu.be/ldVp2fJquq4` and `youtu.be/GoCl7D4asl4`.

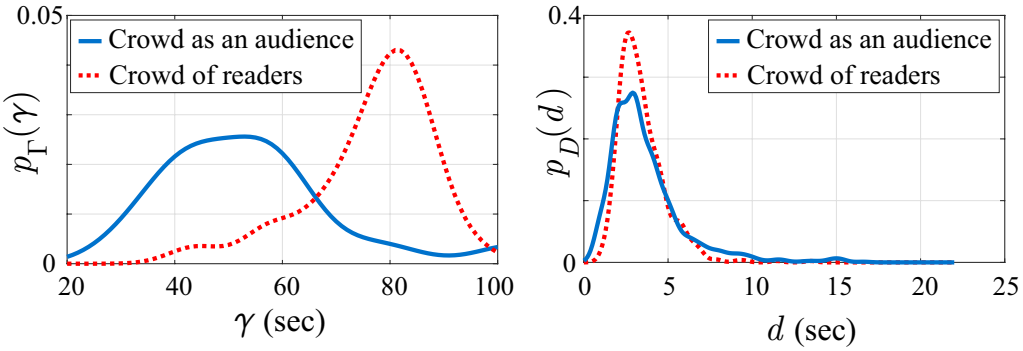


Figure 6.5: (Left) The distribution of the average inter-fidget time of an individual, $p_\Gamma(\gamma)$, for a general population, collected from the data of a total of 30 individuals attending various events in 14 public online Youtube videos (for the case of audience), and 24 individuals reading in 24 different online YouTube videos (for the case of readers). It can be seen that people tend to fidget less when reading, since it requires more focus. (Right) The distribution of the fidget duration $p_D(d)$ of an individual, estimated from the same dataset.

etc), flipping the book's pages, writing a note on the book, or other similar interaction with the reading material, can also be considered as a fidget in this case, and will be registered by the WiFi processing framework as a non-breathing motion.

Fig. 6.5 (left) shows the PDF of the average inter-fidget time of an individual, $p_\Gamma(\gamma)$, in the two aforementioned categories of activities. It can be seen that people tend to fidget less while reading, since the reading activity requires more focus. On the other hand, people tend to fidget once every 50-55 seconds, on average, when attending an event. This result agrees with the findings of Sir. Francis Galton, a famous psychologist in the 19-th century, who observed 50 members of the audience of a public lecture in 1885, and concluded that a person, on average, fidgets once a minute in such gatherings [122]. It also agrees with the findings of the authors of [134], who observed 21 students in a 40-minute lecture, finding that a student, on average, fidgets once every $\sim$57 seconds. Fig. 6.5 (right) then shows the distribution of the fidget duration, $p_D(d)$, of an individual for both categories of activities. The figure suggests that the distribution of the fidget duration is similar for different types of activities, with an average fidget duration of $\sim$2.9

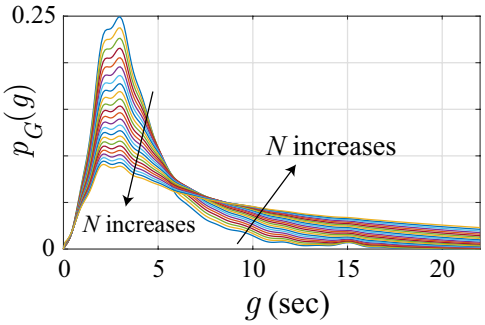


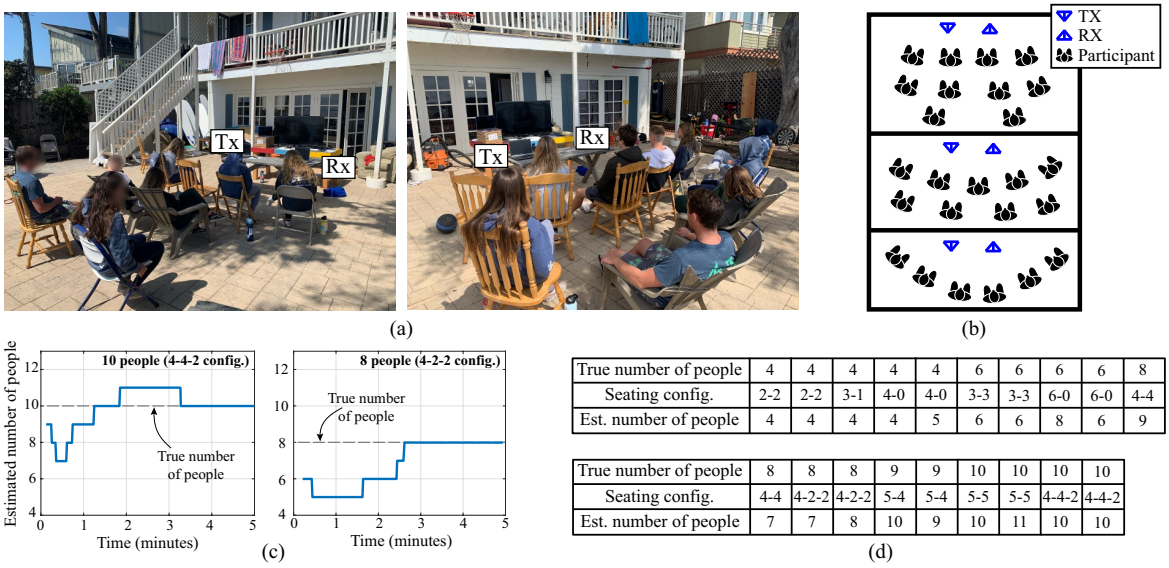
Figure 6.6: PDF of the duration of a CFP of the aggregate process ($p_G(g)$), from Eq. 6.3, as a function of N and for an audience-style activity. The figure explicitly shows the dependence of the duration of a CFP on the number of people, which is characterized through our proposed mathematical framework. Higher N results in a higher probability of having longer CFPs, as can be seen.

seconds.

In order to see how the PDF of CFP (p_G), derived in Eq. 6.3, changes as a function of N, we use the priors of the audience category and plot p_G as a function of N in Fig. 6.6. As can be seen, when N increases, p_G experiences a longer tail, i.e. it is more probable to have longer CFPs, since it is more probable for the fidgets of several people to overlap. Fig. 6.6 further confirms that the WiFi measurements carry the information of the total number of people. Our proposed mathematical framework of Sec. 6.1 then explicitly characterizes this dependency, enabling us to design an end-to-end system to estimate N.

6.4 Experimental Results

In this section, we present the results of our fidget-based crowd counting framework, using WiFi signals, in different environments and with people engaged in different kinds of activities. We start by showing the counting results in non-through-wall environments, in which the Tx and Rx are in the same area as the crowd. Then, we show the counting results in through-wall settings, where the Tx/Rx are placed behind a wall. We further



True number of people	4	4	4	4	4	6	6	6	6	8
Seating config.	2-2	2-2	3-1	4-0	4-0	3-3	3-3	6-0	6-0	4-4
Est. number of people	4	4	4	4	5	6	6	8	6	9

True number of people	8	8	8	9	9	10	10	10	10
Seating config.	4-4	4-2-2	4-2-2	5-4	5-4	5-5	5-5	4-4-2	4-4-2
Est. number of people	7	7	8	10	9	10	11	10	10

(d)

Figure 6.7: (a) Two sample experiments in the first test environment (Area 1): an outdoor patio with several sources of clutter, where up to and including 10 people gather to watch a movie. (b) Seating configuration for three sample experiments in Area 1, where (top) 10 people sit in a 4-4-2 configuration, (middle) 9 people sit in a 5-4 configuration, and (bottom) 6 people sit in a 6-0 configuration. (c) Counting results, as a function of time, for (left) an experiment with 10 people in a 4-4-2 configuration and (right) an experiment with 8 people in a 4-2-2 configuration. (d) Final counting results for all the 19 experiments conducted in Area 1.

show results in different scenarios representing various engagement levels of the crowd, such as attending a lecture, watching a movie, and reading.

6.4.1 Counting in non-through-wall settings

For non-through-wall cases, the Tx and Rx are placed in the same area as the crowd to be counted, with no immediate physical obstruction (such as a wall). However, we emphasize that not all the people in the crowd will have a Line-of-Sight (LoS) to the WiFi link, since they can be blocked by other people/objects. Fig. 6.7 (a) shows two examples of such non-through-wall settings. Overall, we conduct experiments in different environments with people engaged in different categories of activities: audience-style activity and reading activity.

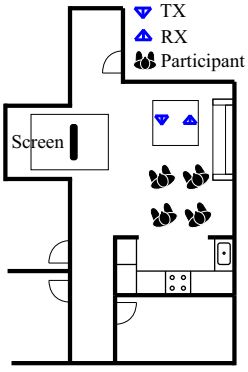
Counting an audience

In this category, we conduct experiments in two different environments, where the crowd is engaged in an audience-style activity, e.g. watching a movie or attending a lecture. Fig. 6.7 (a) shows two sample shots of the first test environment (Area 1), which is an outdoor patio where different number of people (up to and including 10) gather to watch a movie. The patio is cluttered with the walls of the house, multiple furniture items, tables, the TV, some trees, as well as swimming and diving equipment. In this area, we conduct a total of 19 experiments, each lasting for 5 minutes, in which different number of people sit in rows of chairs. We run experiments with several different seating configurations, involving one, two, and three rows. Fig. 6.7 (b) shows the seating configurations for three sample experiments conducted in this area. For example, in the first sample experiment of Fig. 6.7 (b-top), 10 people sit in 3 rows such that 4 people sit in row 1, 4 people in row 2, and 2 people in row 3. We refer to such a seating configuration as 4-4-2. Fig. 6.7 (b-middle), on the other hand, shows a seating configuration where

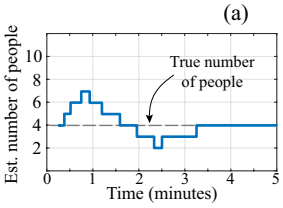
5 people sit in the front row, and 4 people sit in the second row, while Fig. 6.7 (b-bottom) shows a sample configuration where 6 people sit in one row. We refer to any seating configuration by the number of people in each row of chairs (starting from the row closest to the Tx/Rx), separated by hyphens. Fig. 6.7 (c) shows the estimated number of people as a function of time for two sample experiments in this area: 10 people in a 4-4-2 configuration and 8 people in 4-2-2 configuration. Note that the gap at the beginning is due to the fact that our system waits until the end of the first measured CFP, in order to have at least one CFP and one CSP to estimate the number of people using Eq. 6.5. It can be seen that the count estimate starts to converge after about 2 minutes into the data collection. Fig. 6.7 (d) shows the final estimated number of people for all the 19 experiments conducted in this area, involving different number of people and seating configurations. Out of these 19 experiments, our system achieved a counting error of 0 or 1 in 18 experiments (94.74% of the time).

The second test environment (Area 2) for this category is an indoor apartment where a set of $N = 4$ people are watching an online lecture. We conduct a total of 4 experiments in this area, with different seating configurations. Each experiment is 5 minutes long. Fig. 6.8 (a) shows a sample experiment where 4 people are watching a lecture while seated in two rows, with 2 persons in each row. Fig. 6.8 (b) shows the floor plan of the apartment as well as the locations of the participants with respect to the Tx, Rx, and the screen. Fig. 6.8 (c) shows the estimated number of people as a function of time for one sample experiment in which the participants were seated in a 4-0 configuration. It can be seen that the estimated number of people, using our proposed approach, converges to the true number of people. Finally, Fig. 6.8 (d) shows a table of the final counting results for all the 4 experiments in Area 2, where it can be seen that our proposed approach achieves a counting error of 0 or 1 in all 4 experiments, showing a very good counting performance in an indoor environment.

(a)



(b)



(c)

True number of people	4	4	4	4
Seating config.	2-2	2-1-1	4-0	4-0
Est. number of people	3	4	4	5

(d)

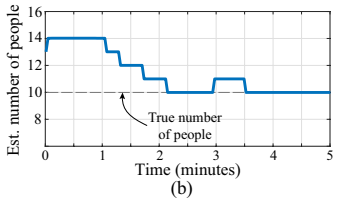
Figure 6.8: (a) The second test environment (Area 2): four residents of an apartment gather in the living room to watch a lecture, in different seating configurations. (b) The floor plan of the apartment showing the locations of the Tx, the Rx, and the 4 participants sitting in a 2-2 configuration. (c) Estimation of the number of people for an experiment where the participants were seated in a 4-0 configuration. (d) The final counting results of the 4 experiments conducted in Area 2, showing a very good counting performance in an indoor environment.

Counting readers

In this category, we conduct 4 experiments, where each person in the crowd is reading. The experiment location (Area 3), shown in Fig. 6.9 (a), is a roofed space (closed from 4 sides) with walls, doors, and vending machines. We conduct four experiments in this area where 8 or 10 people sit in different configurations to read, while a pair of WiFi transceivers are placed in the same area. Fig. 6.9 (b) shows the counting result, as a function of time, for one sample experiment in which 10 people sit in a 4-6 configuration, while Fig. 6.9 (c) shows the final counting results for all the 4 experiments, showing a very good counting performance.

In summary, for non-through-wall scenarios, we conducted a total of 27 experiments in 3 different environments, including an indoor space, a roofed space, and an outdoor space, with different types of activities. Define the counting error, e, as the absolute

(a)



(b)

True number of people	8	8	10	10
Seating Config.	4-4	4-4	4-6	4-6
Est. number of people	7	9	10	10

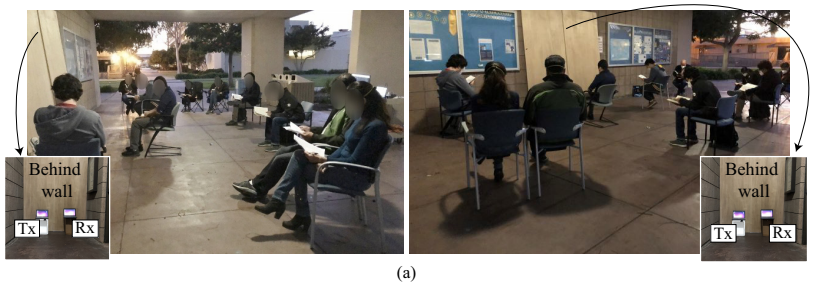
(c)

Figure 6.9: (a) The third test environment (Area 3): a roofed space where 8 or 10 people are engaged in a reading activity. (b) Estimation of the number of people in one sample experiment in which 10 people sit in a 4-6 configuration. (c) The final counting results for all the 4 reading experiments in Area 3, showing a very good counting performance.

difference between the estimated and true number of people ($e = |\hat{N} - N_{\text{true}}|$). We then have an error of 0 or 1 in 26 out of the 27 experiments in non-through-wall settings, i.e. prob($e \leq 1$) = 0.963. Next, define the mean absolute error as MAE = $\mathbb{E}(e)$, and the normalized mean square error as NMSE = $\mathbb{E}\left(\frac{e^2}{N_{\text{true}}^2}\right)$, where $\mathbb{E}(.)$ is the expectation operator. The MAE and NMSE for the non-through-wall experiments are then as follows: MAE = 0.44, and NMSE = 0.015. Finally, the correlation between the true and estimated number of people, defined as $\rho = \frac{\text{Cov}(\hat{N}, N_{\text{true}})}{\sigma_{\hat{N}} \sigma_{N_{\text{true}}}}$, is 0.959, where Cov(.,.) is the covariance between the two variables, and σ is the standard deviation of the corresponding variable. This high correlation coefficient shows that the estimated number of people matches the true number of people well.

6.4.2 Through-wall stationary crowd counting

To further validate our counting framework in more challenging scenarios, we conduct several experiments (with different types of activities) in through-wall scenarios, where the Tx and Rx are placed in a different space outside the seating area (i.e. behind a wall), with no direct LoS to any of the people in the crowd.



(a)

A crowd of readers

True number of people	4	4	4	4	5	5	5	5
Seating config.	2-2	2-2	3-1	3-1	3-2	3-2	3-2	3-2
Est. number of people	5	5	3	4	4	5	4	6

True number of people	6	6	6	6	8	8	10	10
Seating config.	3-3	3-3	3-3	3-3	4-4	4-4	4-6	4-6
Est. number of people	5	5	5	7	7	8	8	9

Crowd as an audience

True number of people	8	8	10	10
Seating config.	4-4	4-4	4-6	4-6
Est. number of people	6	8	10	9

(b)

Figure 6.10: Through-wall crowd counting – (a) Two snapshots of sample experiments in the fourth test environment (Area 4): a roofed space where the Tx/Rx are placed behind a wall. (b) The final counting results for all the 20 experiments in Area 4 involving different activities (reading and watching a movie), showing a very good counting performance in through-wall scenarios.

Counting readers

In this category, we conduct a total of 16 experiments, where we collect WiFi data in Area 4 (two snapshots of which are shown in Fig. 6.10 (a)). Area 4 is a roofed space (closed from 4 sides) in which we place the Tx and Rx behind a wall, as can be seen in the figure. In each of the experiments in this area, different number of people (up to and including 10) sit in different seating configurations while each individual is reading. Fig. 6.10 (b-top) shows the final counting results for all the 16 reading experiments in this area. It can be seen that our proposed approach can achieve a very good counting performance. More specifically, we achieve a counting error of 0 or 1 in 15 out of the 16 experiments (93.75% of the time).

Counting an audience

In this category, we conduct 4 experiments in Area 4 (Fig. 6.10 (a)), where the WiFi Tx and Rx are placed behind a wall. In these experiments, different number of people gather, in different seating configurations, to watch a documentary. Fig. 6.10 (b-bottom) shows the final counting results for these 4 audience-style experiments, showing a very good counting performance.

In summary, for through-wall scenarios, we conducted a total of 20 experiments, with

	Non-through-wall	Through-wall
Number of experiments	27	20
prob(counting error $\leq$ 1)	0.963	0.9
Mean Absolute Error (MAE)	0.44	0.85
Normalized Mean Square Error (NMSE)	0.015	0.028
Correlation coefficient (ρ)	0.959	0.904

Table 6.1: Overall performance of the proposed counting system, over several different areas, activities, and seating configurations.

different types of activities. The counting result was off from the true number of people by 0 or 1 in 18 out of the 20 experiments (i.e. prob($e \leq 1$) = 0.9), with a mean absolute error (MAE) of 0.85, a normalized mean square error (NMSE) of 0.028, and a correlation coefficient (ρ) of 0.904. These results show that our proposed system achieves a very good counting performance even in through-wall cases.

Table 6.1 summarizes the performance over all the areas, activities, and seating configurations for both through-wall and non-through-wall settings.

6.5 Discussions

In this section, we provide a detailed discussion on different aspects of our proposed crowd counting approach, and further motivate future research directions.

Validity of the Poisson model for the fidgeting process

To develop the mathematical model for the fidgeting process in Sec. 6.1, we stated that an individual's fidgeting process is well modeled by a Poisson process, i.e. the inter-fidget times of an individual would follow an exponential distribution. To further validate

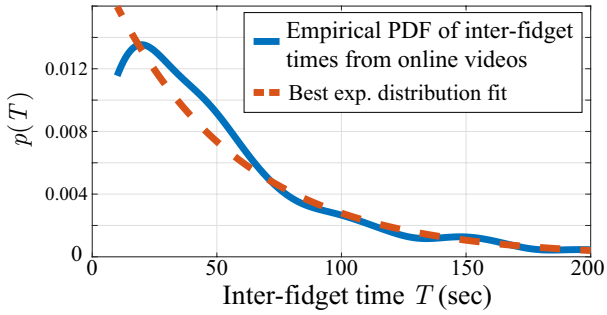


Figure 6.11: The exponential distribution well fits the empirical PDF of the collected individual inter-fidget times of 6 subjects, showing that the Poisson process well describes the individual fidgeting process.

this, Fig. 6.11 plots the empirical PDF of the collected inter-fidget times of 6 people (with very close fidgeting rates so we can put all the data in one pool) from online Youtube videos. The figure also shows the best exponential distribution fit for the empirical data. It can be seen that there is a tight fit between the two, with a very small Kullback-Leibler Distance (KLD) of 0.1834, confirming the validity of the Poisson model.

Robustness to interference by people moving nearby

In real-world scenarios, the stationary crowd can, once in a while, be interrupted by a nearby movement of some walking person, e.g. a person entering/exiting/passing by the area. If the interfering person is close enough to the WiFi transceivers, his/her motion can be picked up by the WiFi receiver. However, such an interference will not affect the counting results much, as long as the rate of these interfering events is not too high. In fact, in several of our experiments (e.g., all those in Areas 3 and 4), pedestrians were passing by frequently, with a rate of 0.3 person per minute (based on observing the pedestrian flow for 90 minutes). However, as all the results of Sec. 6.4 confirmed, we could still count the number of seated people in these areas with a high accuracy.

To further validate this, we conduct two sets of through-wall experiments in Area 4 (shown in Fig. 6.10 (a)). In the first set, we conduct two 5-minute experiments where 5 people sit in the area and read. We then ask an outsider to walk by the WiFi coverage

area frequently and throughout the whole measurement duration, in order to interfere with the WiFi measurements, with an average duration between interfering events of 160 seconds (the interference rate is close to half the average fidgeting rate of an individual). In both experiments, the final counting result was $\hat{N} = 5$, showing that the interference did not affect the counting performance. In the second set of experiments, we conduct two 5-minute experiments where the intruder walks by the WiFi area throughout the whole experiments, with an average duration between interfering events of 90 seconds (the interference rate is close to the average fidgeting rate of an individual). In both these experiments, the final counting result is $\hat{N} = 6$, since the recurring interference events can be considered as the fidgeting of an extra person in the reading group. Overall, if people move by with a rate less than the average fidgeting rate of an individual, their impact will be negligible. Even for higher rates, the impact can be small if the rate is not too high. As part of future work, one can see if the major motion of someone passing by is differentiable from the in-place fidgets in order to properly separate them.

Characterization of the maximum number of people

In this chapter, we have tested our approach in several different settings where the operation area got very crowded, i.e., the density of people per area was high. However, in larger operation areas, even more people can be present and need to be counted. As such, understanding the limitations of a single WiFi link in terms of the maximum number of people it can count is important. Our proposed fidget-based counting system relies on the durations of CSPs/CFPs for counting. As such, the counting performance is expected to degrade when the number of people in the crowd is very large to the extent that there are no CSPs, i.e. at least one person is fidgeting at any point in time, resulting in the aggregate process of Fig. 6.2 becoming a very long fidgeting event. While an exact mathematical characterization of this limitation is part of future work, intuitively, this

situation is more likely to happen when the number of people far exceeds $\tilde{\gamma}/\tilde{d}$, where $\tilde{\gamma}$ is the average inter-fidget time and $\tilde{d}$ is the average fidget duration. As part of future work, one can then utilize more resources, e.g., more links, for counting over larger areas that contain more people.

Robustness to errors in evaluating p_Γ

In this chapter, we have assumed that the general high-level type of an event is known for a given area (e.g., is this a library or a restaurant?), in order to acquire a PDF for p_Γ from available online videos of individuals engaged in similar types of activities (see Sec. 6.3 for more details). While the high-level type of the event would be fairly known for a given area and is thus easily assessable, for the sake of completion, we next study the robustness of the system to the errors in assessing p_Γ. More specifically, instead of using the corresponding priors found for the audience-style and reading-style activities, we set $p_\Gamma(\bar{\gamma}) = 1$ for $\bar{\gamma} = 60$ seconds and zero otherwise, based on the crude characterization of the 1885 paper of Sir. Galton [122]. While the overall counting results expectedly degrade, we find that the counting error is 2 persons or less for 87% of the time across all experiments (with a NMSE of 0.07), which can still be an acceptable performance for many practical applications.

Impact of static multipath

In this chapter, we have tested our proposed approach extensively in indoor, outdoor, and roofed (closed from 4 sides) environments. All these environments experienced a high-level of clutter resulting in static multipath, as we discussed in Section 6.4. Due to the COVID-19 precautionary measures set by our institution, we could not conduct more experiments in indoor settings. However, we anticipate our results to be similar in other indoor areas. This is mainly due to the fact that the static multipath, caused by static objects in the environment, does not affect our system's performance, since its

impact appears in the DC component of the WiFi signals [30], which is removed in the high-pass filtered signal $w^{\text{filt}}(t)$. Our counting results in indoor, roofed, and cluttered outdoor areas further confirm this.

Computational complexity

While the results presented in this chapter are based on the offline processing of the collected CSI data, the design of the algorithm has taken real-time processing considerations into account. For instance, the steps of the WiFi processing pipeline (such as the moving variance calculation) rely on moving windows, which facilitates the implementation of the algorithm in real-time. Moreover, our proposed algorithm is computationally efficient, taking only 11.3 ms, on average, to process one second of the WiFi data, which is smaller than the window size $T_{\text{win}} = 2$ sec, adding no delay to the system when implemented in a real-time manner.

Chapter 7

Conclusions and Future Work

In this dissertation, we proposed novel techniques that enabled new applications for sensing with everyday WiFi signals. In the first part of the dissertation, we showed that given a video footage of a person engaged in some activity, it is possible to generate the corresponding WiFi signals that would have been measured if that person was near a WiFi device. This would then enable new possibilities, two examples of which were shown. More specifically, we showed first that WiFi can identify people walking in an area, from a candidate video footage. Second, we showed that our proposed idea can greatly reduce the training effort of learning-based RF sensing systems, since it can generate an instant RF dataset from the vast available public video dataset, for any set of motion-based human activities. In the second part of the dissertation, we showed that it is possible to use WiFi signals to detect nocturnal seizures in epilepsy patients in a contactless, cheap, fast, and robust manner. Finally, in the last part of the dissertation, we showed that WiFi signals can be used to count the number of people in a stationary (seated) crowd, using the statistics of their aggregate fidgeting behaviors.

We next summarize our results in each of these areas.

7.1 Person Identification From Video Footage

In this part, we proposed XModal-ID, a WiFi-video cross modal person identification system, which can determine if an unknown person walking in a WiFi-covered area is the same as the person in a video footage. To achieve this, XModal-ID utilizes WiFi CSI magnitude measurements of a pair of WiFi transceivers to identify a person, by matching the gait features captured by the WiFi measurements to those from a video of a walking person. XModal-ID does not need any prior wireless or video data of the person to be identified, or the identification area. It can further identify people through walls and does not need the knowledge of the track of the person. In order to evaluate our proposed system, we constructed a large test set with 8 subjects, 5 WiFi areas, and 2 video areas, all of which were unseen in the training phase. Furthermore, the test set includes 3 areas where the transceivers were placed behind a wall, as well as scenarios with complex paths. XModal-ID achieves an overall binary classification accuracy of 85% in predicting whether a WiFi-video pair belong to the same person or not, and top-1, top-2, and top-3 ranking accuracy of 75%, 90%, and 97%, respectively. This demonstrates that our proposed XModal-ID system can robustly identify unknown people in new environments and through walls.

7.2 Multiple People Identification Using Off-the-Shelf WiFi

In this part, we proposed a gait-based identification system that can, for the first time, identify multiple simultaneously walking people through walls, using CSI magnitude measurements of a small number of off-the-shelf WiFi devices. In order to do so, our system first estimates the AoA of the reflected WiFi signals from the walking people, and

uses this information to separate their gait signatures. Given the extracted signal of an individual person's walk, our system generates a spectrogram to capture the frequency-time features of the gait for identification. We have extensively validated our proposed system with 92 test experiments in 4 test areas. In each experiment, 2 or 3 test subjects walk simultaneously in an area. Overall, our system achieves an overall average accuracy of 82% in identifying whether the person of a query data sample (extracted from a multi-person walking experiment) is the same as the person of a candidate spectrogram.

7.3 Teaching RF to Sense Without RF Training Measurements

In this part, we proposed a new and generalizable framework that allows for successfully training RF sensing systems only with already-available video data, and without any real RF data, thus eliminating the traditional labor-intensive phase of collecting real RF training measurements. More specifically, our proposed approach taps into the vast number of available online videos of different human activities/motions, translates them into instant simulated RF data, extracts relevant time-frequency features, and trains a neural network pipeline. Our approach is general and scalable to any motion-based human activity and any given setup. In order to validate our proposed framework, we carried out a case study of gym activity classification using WiFi transceivers. We utilized YouTube videos of the corresponding gym activities, constructed a simulated RF dataset, extracted key features via time-frequency analysis, and trained a classifier, without using any real RF training measurement. After training, the classifier was then extensively tested with real WiFi measurements of 10 subjects performing the 10 gym activities in 3 different test areas. Overall, our system achieved a classification accuracy of 86% when

tested on a small activity period that contains an average of 5.1 repetitions, and 81% when tested on individual repetitions of activities. This demonstrates that the proposed approach can successfully train an RF sensing system with already-available video data, and without any real RF measurements.

7.4 Nocturnal Seizure Detection Using Off-the-Shelf WiFi

In this part, we have considered the problem of nocturnal seizure detection in epilepsy patients using WiFi signals measured on a device placed in the vicinity of the sleeping patient. We first provided a mathematical analysis for the spectral content/bandwidth of the WiFi signal during different kinds of sleep body movements (e.g., seizure, normal movements, and breathing), showing that the bandwidth of the signal can be used to robustly differentiate a seizure from normal movements. We then utilized this analysis to design a robust seizure detection system, which detects all non-breathing body motion events and classifies them, based on their spectral content, to normal movements and seizures. We experimentally validated our proposed system using WiFi CSI data collected from 20 actors in 7 different locations, where they simulated a total of 260 seizures as well as 410 normal sleep movements. Our proposed system achieved a very low probability of false alarm of 0.0097, while being very responsive to seizure events, detecting 93.85% of the seizure instances with an average response time of only 5.69 seconds. These promising results show the potential of using WiFi signals as an accurate and cheap alternative to traditional seizure detection systems

7.5 Stationary Crowd Counting Using Off-the-Shelf WiFi

In this part, we have considered the problem of counting a stationary crowd, using off-the-shelf WiFi. We proposed that the aggregate in-place natural body movements of the crowd carry crucial information on the crowd count. We then developed a mathematical model for the PDFs of the Crowd Fidgeting Periods (CFPs) and Crowd Silent Periods (CSPs) and showed their dependency on the number of people in the area. In developing our mathematical models, we revealed how our problem of interest resembles an old $M/G/\infty$ queuing theory problem, which allowed us to borrow mathematical tools from a 1985 $M/G/\infty$ queuing theory paper. We further showed how to extract the CFPs and CSPs from the received WiFi signal, using the spectral content of the signal. We extensively validated our proposed approach with a total of 47 experiments in four different environments (including both through-wall and non-through-wall settings), in which up to and including $N = 10$ people are seated. We further tested our system with several different number of people, with many different seating configurations, and also with people engaged in a variety of activities. Our evaluation results showed that our proposed approach achieves a very high counting accuracy, with the estimated number of people being only 0 or 1 off from the true number 96.3% of the time in non-through-wall settings, and 90% of the time in through-wall settings.

Material Reuse

The material discussed in the chapters of this dissertation has been published in the following publications:

- Belal Korany, Chitra R. Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage. In the 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19). Association for Computing Machinery, New York, NY, USA, Article 36, 1–15. [30] DOI: https://doi.org/10.1145/3300061.3345437

- Hong Cai, Belal Korany, Chitra R. Karanam, and Yasamin Mostofi. 2020. Teaching RF to Sense without RF Training Measurements. Proceeding of ACM Interactive, Mobile, Wearable Ubiquitous Technology. Vol. 4, No. 4, Article 120. [135] DOI: https://doi.org/10.1145/3432224

- ©IEEE. Reprinted, with permission from Belal Korany, Hong Cai and Yasamin Mostofi. 2021. Multiple People Identification Through Walls Using Off-the-Shelf WiFi. IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6963-6974. [136] DOI: https://doi.org/10.1109/JIOT.2020.3037945.

- Belal Korany and Yasamin Mostofi. 2021. Counting a stationary crowd using off-the-shelf WiFi. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'21). Association for Computing Machinery, New York, NY, USA, 202–214. [137] DOI: https://doi.org/10.1145/3458864.3468012

Bibliography

[1] W. Wang, A. X. Liu, and M. Shahzad, *Gait recognition using WiFi signals*, in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[2] Y. Zeng, P. H. Pathak, and P. Mohapatra, *WiWho: WiFi-based person identification in smart spaces*, in *Proceedings of the International Conference on Information Processing in Sensor Networks*, 2016.

[3] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, *Freesense: Indoor human identification with Wi-Fi signals*, in *Proceedings of the IEEE Global Communications Conference*, 2016.

[4] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, *WiFi-ID: Human identification using WiFi signal*, in *Proceedings of the International Conference on Distributed Computing in Sensor Systems*, 2016.

[5] J. Lv, W. Yang, D. Man, X. Du, M. Yu, and M. Guizani, *Wii: Device-free passive identity identification via WiFi signals*, in *Proceedings of the IEEE Global Communications Conference*, 2017.

[6] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. J. Spanos, *WiFi-based human identification via convex tensor shapelet learning*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[7] S. Depatla and Y. Mostofi, *Crowd counting through walls using wifi*, in *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, IEEE, 2018.

[8] S. Depatla and Y. Mostofi, *Passive crowd speed estimation and head counting using WiFi*, in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9, IEEE, 2018.

[9] Y. Yang, J. Cao, X. Liu, and X. Liu, *Wi-Count: Passing people counting with COTS WiFi devices*, in *IEEE International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9, 2018.

[10] O. T. Ibrahim, W. Gomaa, and M. Youssef, *CrossCount: A Deep learning system for device-free human counting using WiFi*, *IEEE Sensors Journal* **19** (2019), no. 21 9921–9928.

[11] S. Di Domenico, G. Pecoraro, E. Cianca, and M. De Sanctis, *Trained-once device-free crowd counting and occupancy estimation using WiFi: A Doppler spectrum based approach*, in *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 1–8, 2016.

[12] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, *WiFi-enabled smart human dynamics monitoring*, in *Proceedings of the ACM Conference on Embedded Network Sensor Systems*, pp. 1–13, 2017.

[13] H. Zou, Y. Zhou, J. Yang, and C. J. Spanos, *Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT*, *Energy and Buildings* **174** (2018) 309–322.

[14] S. Liu, Y. Zhao, F. Xue, B. Chen, and X. Chen, *DeepCount: Crowd counting with WiFi via deep learning*, *arXiv preprint arXiv:1903.05316* (2019).

[15] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, *Learning sleep stages from radio signals: A conditional adversarial architecture*, in *International Conference on Machine Learning*, pp. 4100–4109, PMLR, 2017.

[16] X. Liu, J. Cao, S. Tang, J. Wen, and P. Guo, *Contactless respiration monitoring via off-the-shelf wifi devices*, *IEEE Transactions on Mobile Computing* **15** (2015), no. 10 2466–2479.

[17] Y. Zhuo, H. Zhu, and H. Xue, *Identifying a new non-linear csi phase measurement error with commodity wifi devices*, in *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 72–79, IEEE, 2016.

[18] J. Wang, Y. Zhao, X. Fan, Q. Gao, X. Ma, and H. Wang, *Device-free identification using intrinsic CSI features*, *IEEE Transactions on Vehicular Technology* **67** (2018), no. 9 8571–8581.

[19] R. Zheng, Y. Zhao, and B. Chen, *Device-free and robust user identification in smart environment using WiFi signal*, in *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing with Applications and the IEEE International Conference on Ubiquitous Computing and Communications*, 2017.

[20] C. Shi, J. Liu, H. Liu, and Y. Chen, *Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT*, in *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2017.

[21] Y. Li and T. Zhu, *Using Wi-Fi signals to characterize human gait for identification and activity monitoring*, in *Proceedings of the IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, 2016.

[22] L. Lee and W. E. L. Grimson, *Gait analysis for recognition and classification*, in *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*, 2002.

[23] D. K. Wagg and M. S. Nixon, *On automated model-based extraction and analysis of gait*, in *Proceedings of the IEEE international conference on Automatic Face and Gesture Recognition*, 2004.

[24] A. Świtoński, A. Polański, and K. Wojciechowski, *Human identification based on gait paths*, in *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*, 2011.

[25] R. T. Collins, R. Gross, and J. Shi, *Silhouette-based human identification from body shape and gait*, in *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*, 2002.

[26] J. Han and B. Bhanu, *Individual recognition using gait energy image*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (2006), no. 2 316–322.

[27] P. Connor and A. Ross, *Biometric recognition by gait: A survey of modalities and features*, *Computer Vision and Image Understanding* **167** (2018) 1–27.

[28] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, *Device-free personalized fitness assistant using WiFi*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2** (2018), no. 4 165.

[29] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, *Understanding and modeling of WiFi signal based human activity recognition*, in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, 2015.

[30] B. Korany, C. R. Karanam, H. Cai, and Y. Mostofi, *XModal-ID: Using WiFi for through-wall person identification from candidate video footage*, in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, 2019.

[31] Y. Ma, G. Zhou, and S. Wang, *WiFi sensing with channel state information: A survey*, *ACM Computing Surveys* **52** (2019), no. 3 46.

[32] Z. Wang, B. Guo, Z. Yu, and X. Zhou, *Wi-Fi CSI-based behavior recognition: From signals and actions to activities*, *IEEE Communications Magazine* **56** (2018), no. 5 109–115.

[33] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, *Device-free WiFi human sensing: From pattern-based to model-based approaches*, *IEEE Communications Magazine* **55** (2017), no. 10 91–97.

[34] Y. Zou, W. Liu, K. Wu, and L. M. Ni, *Wi-Fi radar: Recognizing human behavior with commodity Wi-Fi*, *IEEE Communications Magazine* **55** (2017), no. 10 105–111.

[35] Q. Chen, B. Tan, K. Chetty, and K. Woodbridge, *Activity recognition based on micro-Doppler signature with in-home Wi-Fi*, in *Proceedings of the IEEE International Conference on e-Health Networking, Applications and Services*, 2016.

[36] X. Ma, R. Zhao, X. Liu, H. Kuang, and M. A. A. Al-qaness, *Classification of human motions using micro-Doppler radar in the environments with micro-motion interference*, *Sensors* **19** (2019), no. 11 2598.

[37] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, *Whole-home gesture recognition using wireless signals*, in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, 2013.

[38] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, *Making the invisible visible: Action recognition through walls and occlusions*, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[39] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, *TW-See: Human activity recognition through the wall with commodity Wi-Fi devices*, *IEEE Transactions on Vehicular Technology* **68** (2018), no. 1 306–319.

[40] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, and W. Xu, *Towards environment independent device free human activity recognition*, in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, 2018.

[41] F. Wang, W. Gong, J. Liu, and K. Wu, *Channel selective activity recognition with WiFi: A deep learning approach exploring wideband information*, *IEEE Transactions on Network Science and Engineering* (2018).

[42] F. Xiao, J. Chen, X. Xie, L. Gui, J. Sun, and R. Wang, *SEARE: A system for exercise activity recognition and quality evaluation based on green sensing*, *IEEE Transactions on Emerging Topics in Computing* (2018).

[43] F. Zhang, K. Niu, J. Xiong, B. Jin, T. Gu, Y. Jiang, and D. Zhang, *Towards a diffraction-based sensing approach on human activity recognition*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **3** (2019), no. 1 33.

[44] L. Guo, L. Wang, C. Lin, J. Liu, B. Lu, J. Fang, Z. Liu, Z. Shan, J. Yang, and S. Guo, *Wiar: A public dataset for WiFi-based activity recognition*, *IEEE Access* **7** (2019) 154935–154945.

[45] M. Van der Lende, D. C. Hesdorffer, J. W. Sander, and R. D. Thijs, *Nocturnal supervision and sudep risk at different epilepsy care settings*, *Neurology* **91** (2018), no. 16 e1508–e1518.

[46] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, *Smart homes that monitor breathing and heart rate*, in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 837–846, 2015.

[47] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, *Monitoring vital signs using millimeter wave*, in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 211–220, 2016.

[48] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, *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*, pp. 25–36, 2016.

[49] X. Wang, C. Yang, and S. Mao, *Phasebeat: Exploiting csi phase data for vital sign monitoring with commodity wifi devices*, in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1230–1239, IEEE, 2017.

[50] X. Wang, C. Yang, and S. Mao, *Resbeat: Resilient breathing beats monitoring with realtime bimodal csi data*, in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2017.

[51] F. Wang, F. Zhang, C. Wu, B. Wang, and K. R. Liu, *Respiration tracking for people counting and recognition*, *IEEE Internet of Things Journal* (2020).

[52] T. Nijsen, *Accelerometry based detection of epileptic seizures*, *Book Accelerometry Based Detection of Epileptic Seizures, Series Accelerometry Based Detection of Epileptic Seizures* (2008).

[53] M. Velez, R. S. Fisher, V. Bartlett, and S. Le, *Tracking generalized tonic-clonic seizures with a wrist accelerometer linked to an online database*, *Seizure* **39** (2016) 13–18.

[54] S. Kusmakar, C. K. Karmakar, B. Yan, T. J. O'Brien, R. Muthuganapathy, and M. Palaniswami, *Automated detection of convulsive seizures using a wearable accelerometer device*, *IEEE Transactions on Biomedical Engineering* **66** (2018), no. 2 421–432.

[55] S. Kalitzin, G. Petkov, D. Velis, B. Vledder, and F. L. da Silva, *Automatic segmentation of episodes containing epileptic clonic seizures in video sequences*, *IEEE transactions on biomedical engineering* **59** (2012), no. 12 3379–3385.

[56] E. E. Geertsema, R. D. Thijs, T. Gutter, B. Vledder, J. B. Arends, F. S. Leijten, G. H. Visser, and S. N. Kalitzin, *Automated video-based detection of nocturnal convulsive seizures in a residential care setting*, *Epilepsia* **59** (2018) 53–60.

[57] S. Beniczky, T. Polster, T. Kjaer, and H. Hjalgrim, *Detection of generalized tonic–clonic seizures by a wireless wrist accelerometer: a prospective, multicenter study*, *Epilepsia* **54** (2013), no. 4 e58–e61.

[58] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. Pearl, and T. Loddenkemper, *Automated seizure detection systems and their effectiveness for each type of seizure*, *Seizure* **40** (2016) 88–101.

[59] Medpage, *Mp5 seizure movement detection monitor*, 2020.

[60] C. Chen, Y. Han, Y. Chen, H.-Q. Lai, F. Zhang, B. Wang, and K. R. Liu, *TR-BREATH: Time-reversal breathing rate estimation and detection*, *IEEE Transactions on Biomedical Engineering* **65** (2017), no. 3 489–501.

[61] Y.-K. Cheng and R. Y. Chang, *Device-free indoor people counting using WiFi channel state information for Internet of Things*, in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2017.

[62] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, *End-to-end recovery of human shape and pose*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[63] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[64] W. C. Chew, *Waves and fields in inhomogeneous media*. IEEE press, 1995.

[65] S. Katz, A. Tal, and R. Basri, *Direct visibility of point sets*, *ACM Transactions on Graphics* **26** (2007), no. 3 24.

[66] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, *Capturing the human figure through a wall*, *ACM Transactions on Graphics* **34** (2015), no. 6 219.

[67] C. R. Karanam, B. Korany, and Y. Mostofi, *Magnitude-based angle-of-arrival estimation, localization, and target tracking*, in *Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2018.

[68] I. Orović, S. Stanković, and M. Amin, *A new approach for classification of human gait based on time-frequency feature representations*, *Signal Processing* **91** (2011), no. 6 1448–1456.

[69] A. Seifert, M. Amin, and A. M. Zoubir, *Toward unobtrusive in-home gait analysis based on radar micro-Doppler signatures*, *IEEE Transactions on Biomedical Engineering* (2019).

[70] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, *Tool release: Gathering 802.11n traces with channel state information*, *ACM SIGCOMM Computer Communication Review* **41** (2011), no. 1 53–53.

[71] A. Sinha, R. Gupta, and A. Kutnar, *Sustainable development and green buildings*, *Drvna Industrija* **64** (2013), no. 1 45–53.

[72] J. L. Geisheimer, E. F. Greneker, and W. S. Marshall, *High-resolution Doppler model of the human gait*, in *Radar Sensor Technology and Data Visualization*, 2002.

[73] N. V. Chawla, *Data mining for imbalanced datasets: An overview*, in *Data mining and knowledge discovery handbook*, pp. 875–886. Springer, 2009.

[74] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, *Automatic differentiation in PyTorch*, .

[75] M. Franěk, *Environmental factors influencing pedestrian walking speed*, *Perceptual and Motor Skills* **116** (2013), no. 3 992–1019.

[76] K. He, G. Gkioxari, P. Dollar, and R. Girshick, *Mask R-CNN*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[77] C. R. Karanam, B. Korany, and Y. Mostofi, *Tracking from one side: Multi-person passive tracking with WiFi magnitude measurements*, in *Proceedings of the International Conference on Information Processing in Sensor Networks*, 2019.

[78] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, *Sonar tracking of multiple targets using joint probabilistic data association*, *IEEE Journal of Oceanic Engineering* **8** (1983), no. 3 173–184.

[79] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, *Tool release: Gathering 802.11n traces with channel state information*, *ACM SIGCOMM Computer Communication Review* **41** (2011), no. 1 53–53.

[80] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, *MPCA: Multilinear principal component analysis of tensor objects*, *IEEE Transactions on Neural Networks* **19** (2008), no. 1 18–39.

[81] R. Wilson, *Propagation losses through common building materials 2.4 GHz vs. 5 GHz*, tech. rep., University of Southern California, 2002.

[82] K. P. Murphy, *Machine learning: A probabilistic perspective*. MIT Press, 2012.

[83] D. Gafurov, E. Snekkenes, and P. Bours, *Spoof attacks on gait authentication system*, *IEEE Transactions on Information Forensics and Security* **2** (2007), no. 3 491–502.

[84] C.-C. Yu, *Sidelobe reduction of asymmetric linear array by spacing perturbation*, *Electronics Letters* **33** (1997), no. 9 730–732.

[85] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, *Neural body fitting: Unifying deep learning and model based human pose and shape estimation*, in *Proceedings of the International Conference on 3D Vision*, 2018.

[86] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, *Human mesh recovery from monocular images via a skeleton-disentangled representation*, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[87] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, *Bodynet: Volumetric inference of 3D human body shapes*, in *Proceedings of the European Conference on Computer Vision*, 2018.

[88] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, *SMPL: A skinned multi-person linear model*, *ACM Transactions on Graphics* **34** (2015), no. 6 248.

[89] S. Depatla, L. Buckland, and Y. Mostofi, *X-ray vision with only WiFi power measurements using rytov wave models*, *IEEE Transactions on Vehicular Technology* **64** (2015), no. 4 1376–1387.

[90] W. C. Gibson, *The method of moments in electromagnetics*. Chapman and Hall/CRC, 2014.

[91] J.-M. Jin, *The finite element method in electromagnetics*. John Wiley & Sons, 2015.

[92] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, *Rethinking the Faster R-CNN architecture for temporal action localization*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[93] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, *End-to-end learning of action detection from frame glimpses in videos*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[94] T. Zhang, B. Huang, and Y. Wang, *Object-occluded human shape and pose estimation from a single color image*, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2020.

[95] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask R-CNN*, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[96] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, *IndoTrack: Device-free indoor human tracking with commodity Wi-Fi*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1** (2017), no. 3 72.

[97] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, *Widar2.0: Passive human tracking with a single Wi-Fi link*, in *Proceedings of the 1International Conference on Mobile Systems, Applications, and Services*, 2018.

[98] J. Wang, H. Jiang, J. Xiong, K. Jamieson, X. Chen, D. Fang, and B. Xie, *LiFS: Low human-effort, device-free localization with fine-grained subcarrier information*, in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, 2016.

[99] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, *Zero-effort cross-domain gesture recognition with Wi-Fi*, in *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, 2019.

[100] Y. Xie, J. Xiong, M. Li, and K. Jamieson, *mD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking*, in *Proceedings of the International Conference on Mobile Computing and Networking*, 2019.

[101] S. D. Lhatoo, M. Nei, M. Raghavan, M. Sperling, B. Zonjy, N. Lacuey, and O. Devinsky, *Nonseizure sudep: sudden unexpected death in epilepsy without preceding epileptic seizures*, *Epilepsia* **57** (2016), no. 7 1161–1168.

[102] S. Jenssen, E. J. Gracely, and M. R. Sperling, *How long do most seizures last? a systematic comparison of seizures recorded in the epilepsy monitoring unit*, *Epilepsia* **47** (2006), no. 9 1499–1503.

[103] D. G. Tucker, *The invention of frequency modulation in 1902*, *The Radio and Electronic Engineer* **40** (1970) 33–37.

[104] B. P. Lathi, *Modern Digital and Analog Communication Systems*. Oxford University Press, Inc., 1998.

[105] J. R. Carson, *Notes on the theory of modulation*, *Proceedings of the Institute of Radio Engineers* **10** (1922), no. 1 57–64.

[106] J. Zhang, W. Xu, W. Hu, and S. Kanhere, *Wicare: Towards in-situ breath monitoring*, in *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 126–135, 2017.

[107] K. Barrett, S. Barman, H. Brooks, and J. Yuan, *Ganong's review of medical physiology*. McGraw-Hill Education, 2019.

[108] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, *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*, pp. 267–276, 2015.

[109] R. Q. Quiroga, H. Garcia, and A. Rabinowicz, *Frequency evolution during tonic-clonic seizures*, *Electromyography and clinical neurophysiology* **42** (2002), no. 6 323–332.

[110] H. Lüders, J. Acharya, C. Baumgartner, S. Benbadis, A. Bleasel, R. Burgess, D. Dinner, A. Ebner, N. Foldvary, E. Geller, *et. al.*, *Semiological seizure classification*, *Epilepsia* **39** (1998), no. 9 1006–1013.

[111] K. Mahr, M. Bergmann, L. Kay, L. Möller, P. Reif, L. Willems, K. Menzler, S. Schubert-Bast, K. Klein, S. Knake, *et. al.*, *Prone, lateral, or supine positioning at seizure onset determines the postictal body position: A multicenter video-eeg monitoring cohort study*, *Seizure* **76** (2020) 173–178.

[112] J. De Koninck, D. Lorrain, and P. Gagnon, *Sleep positions and position shifts in five age groups: an ontogenetic picture*, *Sleep* **15** (1992), no. 2 143–149.

[113] S. Coussens, M. Baumert, M. Kohler, J. Martin, D. Kennedy, K. Lushington, D. Saint, and Y. Pamula, *Movement distribution: a new measure of sleep fragmentation in children with upper airway obstruction*, *Sleep* **37** (2014), no. 12 2025–2034.

[114] T. Nijsen, R. M. Aarts, P. Cluitmans, and P. Griep, *Time-frequency analysis of accelerometry data for detection of myoclonic seizures*, *IEEE Transactions on Information Technology in Biomedicine* **14** (2010), no. 5 1197–1203.

[115] O. Walch, Y. Huang, D. Forger, and C. Goldstein, *Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device*, *Sleep* **42** (2019), no. 12 zsz180.

[116] H. S. Barrows *et. al.*, *An overview of the uses of standardized patients for teaching and evaluating clinical skills*, *Academic Medicine-Philadelphia-* **68** (1993) 443–443.

[117] B. A. Dworetzky, S. Peyre, E. J. Bubrick, T. A. Milligan, S. J. Yule, H. Doucette, and C. N. Pozner, *Interprofessional simulation to improve safety in the epilepsy monitoring unit*, *Epilepsy & Behavior* **45** (2015) 229–233.

[118] J. van Andel, R. Thijs, A. de Weerd, J. Arends, and F. Leijten, *Non-eeg based ambulatory seizure detection designed for home use: What is available and how will it influence epilepsy care?*, *Epilepsy & Behavior* **57** (2016) 82–89.

[119] U. Kramer, S. Kipervasser, A. Shlitner, and R. Kuzniecky, *A novel portable seizure detection alarm system: preliminary results*, *Journal of Clinical Neurophysiology* **28** (2011), no. 1 36–38.

[120] K. V. Poppel, S. P. Fulton, A. McGregor, M. Ellis, A. Patters, and J. Wheless, *Prospective study of the emfit movement monitor*, *Journal of child neurology* **28** (2013), no. 11 1434–1436.

[121] J. Arends, R. Thijs, T. Gutter, C. Ungureanu, P. Cluitmans, J. Van Dijk, J. van Andel, F. Tan, A. de Weerd, B. Vledder, *et. al.*, *Multimodal nocturnal seizure detection in a residential care setting: A long-term prospective trial*, *Neurology* **91** (2018), no. 21 e2010–e2019.

[122] F. Galton, *The measure of fidget*, *Nature* **32** (1885), no. 817 174–175.

[123] W. Stadje, *The busy period of the queuing system $M/G/\infty$*, *Journal of Applied Probability* (1985) 697–704.

[124] G. Last and M. Penrose, *Lectures on the Poisson process*, vol. 7. Cambridge University Press, 2017.

[125] D. Lin, E. Grimson, and J. W. Fisher, *Construction of dependent Dirichlet processes based on Poisson processes*, in *Advances in neural information processing systems*, pp. 1396–1404, 2010.

[126] V. V. Kalashnikov, *Mathematical methods in queuing theory*, vol. 271. Springer Science & Business Media, 2013.

[127] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[128] B. Korany and Y. Mostofi, *Nocturnal seizure detection using off-the-shelf WiFi*, *arXiv preprint arXiv:2103.13556* (2021).

[129] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, *Widar: Decimeter-level passive tracking via velocity monitoring with commodity WiFi*, in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 1–10, 2017.

[130] R. H. Venkatnarayan, G. Page, and M. Shahzad, *Multi-user gesture recognition using WiFi*, in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 401–413, 2018.

[131] Y. Nam, Y. Kong, B. Reyes, N. Reljin, and K. H. Chon, *Monitoring of heart and breathing rates using dual cameras on a smartphone*, *PloS one* **11** (2016), no. 3 e0151013.

[132] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, *A non-EEG biosignals dataset for assessment and visualization of neurological status*, in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 110–114, IEEE, 2016.

[133] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals*, *Circulation* **101** (2000 (June 13)), no. 23 e215–e220.

[134] J. Farley, E. Risko, and A. Kingstone, *Everyday attention and lecture retention: The effects of time, fidgeting, and mind wandering*, *Frontiers in psychology* **4** (2013) 619.

[135] H. Cai, B. Korany, C. R. Karanam, and Y. Mostofi, *Teaching rf to sense without rf training measurements*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **4** (2020), no. 4 1–22.

[136] B. Korany, H. Cai, and Y. Mostofi, *Multiple People Identification Through Walls Using Off-the-Shelf WiFi*, *IEEE Internet of Things Journal* **8** (2021), no. 8 6963–6974.

[137] B. Korany and Y. Mostofi, *Counting a Stationary Crowd Using Off-the-Shelf WiFi*, in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 202–214, 2021.