1222·2022
800 ANNI

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITY OF PADOVA

DEPARTMENT OF INFORMATION ENGINEERING
Ph.D. Course in Information Engineering
Information and Communication Science and Technologies Curriculum
XXXV series

Human Sensing with mmWave Systems: from RADAR to Integrated Sensing and Communication

Ph.D. Candidate
Jacopo Pegoraro

Ph.D. Supervisor
Professor Michele Rossi

Ph.D. Coordinator
Professor Andrea Neviani

Academic Year
2021–2022

Abstract

Contactless perception of human activity holds the potential to revolutionize the ways we interact with technology and our surroundings, enabling completely new remote, unobtrusive monitoring systems. In this context, the use of Millimeter-Wave (mmWave) reflected radio signals to detect, track, and analyze the movement of people leveraging the Radio Detection and Ranging (RADAR) principle has sparked great interest from academia and the industry alike. This is motivated by the high sensitivity and robustness of such frequencies in perceiving and identifying small-scale movement of the human body parts while being less privacy-invasive than widely adopted camera systems as no visual representation of the scene is captured.

However, despite its promising features, mmWave human sensing poses several challenges. The high sensitivity of mmWaves makes the mathematical modeling of the reflections on the human body extremely complex, while the high attenuation occurring at such frequencies raises the question of what kind of transceivers should be used, how to deploy them to provide good coverage, and how to combine the obtained information with other sensors. Leveraging the channel estimation process of wireless communication devices to endow them with RADAR-like capabilities holds great potential to solve these problems. Future wireless networks are expected to be extremely dense, with billions of connected devices continuously exchanging signals which could be *reused* to obtain information on the surroundings at almost zero cost.

This thesis makes substantial contributions to the field of mmWave human sensing by advancing the state-of-the-art along two research lines. *First*, we focus on pure sensing, exploring the potential of dedicated mmWave RADAR devices for indoor people tracking and *identification*. We develop algorithms that can exploit the reflected signal properties to obtain the position in space of multiple subjects, and extract Doppler-related features of their gait (i.e., their individual way of walking) to recognize their identities. Then, we utilize such algorithms to solve the important and timely problem of unobtrusive crowd monitoring in indoor environments, proposing a sensor fusion method to combine thermal images with mmWave RADAR gait signatures. *Second*, we leverage mmWave RADAR signal processing methods to address Integrated Sensing And Communication (ISAC), proposing the first approach to retrofit next-generation mmWave Wi-Fi Access Points (APs) into multipurpose devices that, in addition to providing connectivity, can also detect, track, and recognize the movements of people in their surroundings. To this end, we leverage the properties of the mmWave channel to reconstruct human movement features from irregular and sparse communication packets, thus fully reusing them for sensing purposes.

The methodology adopted in this thesis is to integrate and jointly develop standard signal processing techniques and data-driven machine learning algorithms. Our claims are backed by extensive on-field experimentation with cutting-edge mmWave RADAR and ISAC research testbeds. This approach represents the most promising way to develop future mmWave sensing systems and to achieve the envisioned goal of pervasive, human-oriented remote perception technology.

Contents

Listing of figures

Listing of tables

Listing of acronyms

Symbols

μD micro-Doppler

3GPP 3$^{\text{rd}}$ Generation Partnership Project

4G Fourth Generation

5G Fifth Generation

6G Sixth Generation

A

AoA Angle of Arrival

AP Access Point

B

BP Beam Pattern

C

CA-CFAR Cell-Averaging Constant False Alarm Rate

CDF Cumulative Distribution Function

CE Cross-Entropy

CEF Channel Estimation Field

CFAR Constant False Alarm Rate

CFO Carrier Frequency Offset

CFR Channel Frequency Response

CIR Channel Impulse Response

CJPDA Cheap Joint Probabilistic Data Association

CM-KF Converted-Measurements Kalman Filter

CNN Convolutional Neural Network

COTS Commercial-Off-The-Shelf

CS Compressive Sensing

CSI Channel State Information

CSI-RS CSI-Reference Signal

CV Constant Velocity

D

DBSCAN Density-Based Spatial Clustering for Applications with Noise

DFT Discrete Fourier Transform

DL Deep Learning

DMG Directional Multi Gigabit

E

EKF Extended Kalman Filter

ELM Extreme Learning Machines

ELU Exponential-Linear Unit

EOT Extended Object Tracking

F

FC Fully-Connected

FMCW Frequency-Modulated Continuous-Wave

FoV Field-of-View

FPA Focal Plane Array

FPGA Field Programmable Gate Array

FT Fourier Transform

G

GM Gaussian Mixture

GPU Graphical Processing Unit

GRU Gated Recurrent Unit

GT Ground Truth

H

HAR Human Activity Recognition

HT High Throughput

I

IF Intermediate Frequency

IFS Inter-Frame Spacing

IHT Iterative Hard Thresholding

ISAC Integrated Sensing And Communication

J

JPDAF Joint Probabilistic Data Association Filter

K

KF Kalman Filter

L

LOS Line-of-Sight

LSTM Long Short-Term Memory

M

MCS Modulation and Coding Scheme

MIMO Multiple Input Multiple Output

ML Machine Learning

MLP Multi-Layer Perceptron

mmWave Millimeter-Wave

MOTA Multiple Object Tracking Accuracy

MTI Moving Target Indication

MTT Multi-Target Tracking

N

NETD Noise Equivalent Temperature Difference

NN Neural Network

NN-CJPDA Nearest-Neighbors Cheap Joint Probabilistic Data Association

NN-JPDA Nearest-Neighbors Joint Probabilistic Data Association

NR New Radio

O

OFDM Orthogonal Frequency Division Multiplexing

P

PC Point-Cloud

PDU Protocol Data Unit

PHY Physical Layer

R

RADAR Radio Detection and Ranging

RCS Radar Cross-Section

RD Range-Doppler

RDA Range-Doppler-Azimuth

RF Radio Frequency

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

ROI Region Of Interest

RSSI Received Signal Strength Indication

S

SC Single Carrier

SDR Software-Defined Radio

SLS Sector Level Sweep

SSB Synchronization Signal Block

STF Short Training Field

STFT Short Time Fourier Transform

T

TC Thermal Camera

TConv Temporal Convolution

TCPCN Temporal Convolution Point-Cloud Network

TDM Time-Division Multiplexing

TF Time-Frequency

U

UART Universal Asynchronous Receiver-Transmitter

USB Universal Serial Bus

UWB Ultra Wide-Band

V

VHT Very High Throughput

W

WELM Weighted Extreme Learning Machines

WLAN Wireless Local Area Network

Y

YOLOv3 You Only Look Once v3

What a scholar one might be if one knew well only some half a dozen books.

Gustave Flaubert

1 Introduction

Contactless perception of the surroundings is a key human capability. We heavily rely on our visual system to gather information about the environment and other people *at distance*, to obtain actionable insights to guide our decisions. This has been reproduced to some extent in camera-based sensor systems, which are nowadays a fundamental building block of a wide range of technologies. However, visual systems are limited to capturing the optical spectrum of electromagnetic waves, and thus are subject to performance degradation in the dark, with adverse weather conditions, or in presence of smoke.

For these reasons, in several applications other sensors have been developed that can remotely sense the environment using different frequency ranges, e.g., microwaves, which span frequencies from 300 MHz to 300 GHz. This is the case of Radio Detection and Ranging (RADAR), which has been used for several decades in military and civilian applications for detecting and tracking targets of interest. Thanks to the longer wavelength of microwaves with respect to visible light, RADAR devices work independently of lighting conditions and are regarded as *all-weather* sensors, meaning they do not suffer from significant performance degradation in adverse weather conditions.

In recent years, RADAR has found novel application in the form of highly accurate, short-range sensors for indoor and outdoor human movement sensing. This is due to recent development in Radio Frequency (RF) transmitter-receivers (*transceivers*) along two lines. On the one hand, in the wireless communication networks field, efficient RF front-ends have been designed to transmit in the Millimeter-Wave (mmWave) frequency range (30-300 GHz). The mmWave band includes much higher frequencies than those traditionally used for communication, thus granting available space for wideband channels capable of supporting the ever-increasing demand for higher data rates. On the other hand, antennas working at mmWaves can a have much smaller form factor, as this is intrinsically related to the wavelength of the transmitted signal, thus allowing the design of Multiple Input Multiple Output (MIMO) transceivers featuring large *antenna arrays*. In wireless communication, these technological advances represent disruptive elements that have put forward

3$^{\text{rd}}$ Generation Partnership Project (3GPP) Fifth Generation (5G) cellular networks and the so-called Wi-Gig standards IEEE 802.11ad/ay for Gigabit Wireless Local Area Networks (WLANs).

From the RADAR perspective, the combination of mmWave and MIMO allows *(i)* unprecedented accuracy in measuring distances between objects (*ranging*), *(ii)* high sensitivity to the Doppler effect, and *(iii)* the capability of measuring the Angle of Arrival (AoA) of the signal reflections, which can be used to infer the spatial location of the targets. Point *(i)* stems from the fact that the RADAR ranging accuracy improves by using wider signal bandwidth, of which there is large availability at mmWaves, while point *(ii)* is due to the usage of high carrier frequencies. Point *(iii)* instead derives from the spatial diversity granted by using MIMO antenna arrays, as small signal phase differences at the different antennas can be used to compute the direction of arrival of the incoming waveform. Thanks to these properties, in the past few years compact and low-cost MIMO mmWave RADAR devices have been produced and used by researchers in the most diverse highly accurate human sensing applications, such as fine-grained tracking of respiratory activity and heartbeat [1], advanced gesture recognition [2]–[4], and gait-based person identity recognition [5]–[7].

In the first part of this thesis, consisting of Chapter 2 and Chapter 3, we focus on the latter problem of exploiting RADAR reflections to distinguish between different individuals based on their gait. This could have potentially disruptive applications in surveillance systems, individually-tailored smart home services, and automated patient monitoring in hospitals or remote healthcare. Gait features are encoded in the RADAR return signal as it is reflected back from a person's body during walking. Indeed, the small-scale micro-Doppler (μD) effect caused by the movement of the limbs induces a detectable frequency modulation on the reflection. However, reliably extracting these features from highly noisy and cluttered mmWave signals using standard mathematical analysis is infeasible, as available tractable models oversimplify the underlying human movement patterns and the complex mmWave propagation. To solve this challenge, in Chapter 2 we propose tightly integrated RADAR signal processing and machine learning algorithms that can detect, track, and identify multiple subjects concurrently moving in an indoor space. In this sense, our main contribution is showing that, on the one hand, tracking the movement trajectories of subjects helps in separating their movement features. On the other hand, identifying subjects can be useful to correct mistakes of the tracking process, enhancing its reliability and accuracy. In Chapter 2 we move beyond the person identification task, and we investigate the integration of mmWave RADAR and infrared thermal imaging, developing new sensor fusion approaches. Our aim is to design a joint body temperature screening and interpersonal distance monitoring method for preventing the spread of contagious diseases. In this respect, we leverage the mmWave RADAR people tracking capabilities to measure distances between people in real-time. At the same time, their gait features are extracted and used to perform contact tracing across different rooms by recognizing subjects from their way of walking. This is done using a novel combination of learning algorithms that work on-the-fly, on subjects that were not seen during the training phase of the system. In addition, a new sensor fusion algorithm to associate the temperature readings with the correct RADAR targets is presented.

Despite providing promising results in human sensing tasks, mmWave MIMO RADAR also

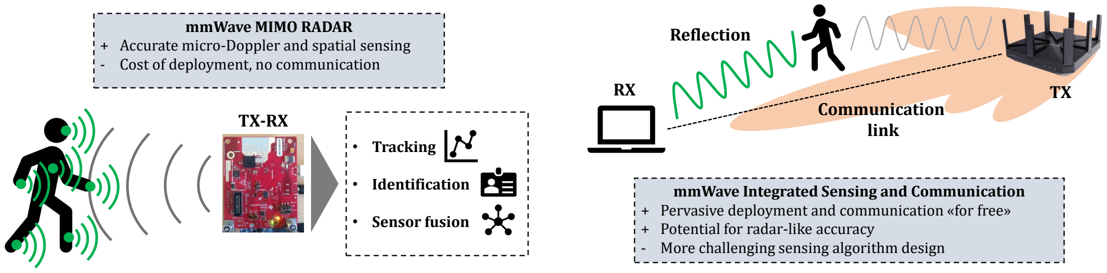

mmWave MIMO RADAR
+ Accurate micro-Doppler and spatial sensing
- Cost of deployment, no communication
TX-RX
Tracking
Identification
Sensor fusion
Reflection
RX
TX
Communication link
mmWave Integrated Sensing and Communication
+ Pervasive deployment and communication «for free»
+ Potential for radar-like accuracy
- More challenging sensing algorithm design

Figure 1.1: The two human sensing paradigms studied in this thesis: mmWave MIMO RADAR (left) and mmWave ISAC (right).

presents evident drawbacks in terms of cost and ease of deployment. Commercial devices have a limited range (up to 6-8 m) [8] and are subject to occlusion due to the limited penetration capabilities of mmWaves. Covering medium to-large indoor spaces thus requires multiple, networked RADAR sensors, increasing the overall deployment cost and complexity. Moreover, RADAR devices are dedicated sensors, meaning they entail spectrum occupancy and energy consumption for the sole sensing purpose. This limitation, together with the ubiquitous deployment of Wi-Fi and cellular communication devices have sparked research interest towards developing Integrated Sensing And Communication (ISAC) technology, to avoid the cost of installing dedicated hardware while at the same time benefiting from communication capabilities. This recent trend has led to the identification of sensing as a key feature of next generation Sixth Generation (6G) mobile networks [9] and the creation of the IEEE 802.11bf standardization group [10], aimed at integrating sensing techniques into Wi-Fi Access Points (APs). In the second part of the thesis (Chapter 4 and Chapter 5) we address the ISAC problem, proposing ways of repurposing communication devices by endowing them with additional environment sensing capabilities. Our focus is on the *Wi-Gig* technology, as standardized by the IEEE 802.11ay group, which works in the unlicensed 60 GHz band of the mmWave spectrum. This technology represents the next generation of Wi-Fi systems, targeting high bitrate applications in WLANs, such as augmented or virtual reality. First, in Chapter 4, we demonstrate how it is possible to perform simultaneous tracking, activity recognition, and identification of people concurrently moving in a room, using the reflections of standard-compliant waveforms used for channel estimation. We do so by proposing the first approach to extract μD signatures from IEEE 802.11ay packets, validating our method on real measurements obtained with a Software-Defined Radio (SDR) testbed. Then, in Chapter 5, we address an even deeper integration of sensing and communication, in which the underlying communication traffic is assumed to be sparse and irregular as in real Wi-Fi traces. This poses several challenges for the μD extraction, that we solve by introducing a novel sparse reconstruction method based on the intrinsic sparsity of the mmWave channel. Our approach effectively reduces both the overhead and the channel occupation caused by the additional sensing task by several times, while obtaining the same, or even better, sensing accuracy. The mmWave RADAR human sensing and ISAC fields are schematized in Fig. 1.1 and detailed in the remaining part of this introduction (Section 1.1 and Section 1.2). This will briefly present the necessary background material for the technical discussion given in the following chapters.

1.1 Sensing with mmWave MIMO RADAR devices

The basic RADAR working principle is to transmit a microwave pulse and collect the signal copies that are reflected back by obstacles in the environment. Typically, the receiver applies amplification, down-conversion, and analog-to-digital conversion to enable subsequent processing. In a second phase, signal processing algorithms are applied to separate the desired targets from the spurious reflections coming from background objects, termed *clutter*. Next, the estimation of distance (*range*), velocity, angular position, and other properties of the targets is carried out. A key feature of RADAR is its ability to measure the *Doppler effect*, which causes a frequency shift in the reflected radio waves due to the movement of the target in space. This is exploited in many RADAR systems for the estimation of the velocity of the target, as this is proportional to the frequency shift. In the following, we introduce Frequency-Modulated Continuous-Wave (FMCW) RADAR systems, which will be used extensively in the first part of the thesis, and the person identification problem based on gait features extracted from RADAR reflections.

Frequency Modulated Continuous Wave RADAR

Many RADAR systems that enable the joint estimation of range and Doppler effect of the target follow the so-called *pulse-Doppler* principle. The transmitted signal is a short sinusoidal pulse, which allows computing the range measuring the time needed for the reflection to reach the receiver, and obtaining the velocity of the target from the measured frequency shift. However, such systems are often expensive and difficult to manufacture as they have a high peak-to-average power ratio, and they require costly circuitry to accurately measure very short return time differences. To provide cheap RADAR devices to be deployed indoors and outdoors for human movement sensing, these drawbacks have to be solved. One possible way of doing so is to use FMCW systems. In this kind of RADAR, short sinusoidal pulses are replaced by longer *chirp* waveforms, whose frequency is linearly swept over a pre-defined interval. This mitigates the peak-to-average power ratio problem and enables measuring range from the frequency difference between the transmitted and the received signals rather than from time differences, lifting the high accuracy requirements on the receiver circuitry. For these reasons FMCW RADAR devices have been widely used in recent indoor monitoring and automotive applications [5], [11], [12]. Commercial devices can use a chirp bandwidth of up to $4-5$ GHz, reaching centimeter-level ranging accuracy.

FMCW chirps are typically transmitted in *frames* of L elements, followed by a waiting period. We refer to the chirp repetition period inside a frame as T, while the frame repetition period is denoted by T_{rep}. At the receiver, the incoming signal is *mixed* (i.e., multiplied) with the transmitted one yielding a narrowband Intermediate Frequency (IF) signal from which the targets' parameters can be estimated. After sampling, the IF signals coming from the different chirps are arranged in a two-dimensional matrix in which samples from the same chirp are arranged as column vectors and stacked together. The resulting matrix has dimension $N \times L$, where N is the number of samples taken from a single chirp. In a MIMO RADAR system, such processing is performed at each of the M receiver antennas, resulting in a so-called *RADAR cube* of dimension $N \times L \times M$. As described in detail in Chapter 2, the target properties of interest can be extracted via 3D spectrum analysis of the RADAR cube.

micro-Doppler signatures

In this thesis we focus on new applications of RADAR that touch on our day-to-day living, involving the study of how human movement features are embedded in RADAR return signals. A large body of research in the last few years has shown how these features can be accurately reconstructed from the micro-Doppler (μD) signature of the movement, computed through spectral analysis of the reflected signal. The μD concept was introduced in the seminal work of Chen [13], [14], that showed how targets involving multiple moving parts cause a complex Doppler frequency *modulation* on the waveform, rather than a simple shift. In humans different body parts, each with *its own velocity*, are involved in all common daily activities, like walking, running, sitting down, etc. This is reflected in the μD modulation of the RADAR reflection, which is different for different activities, individuals, and gestures. To enable accurate analysis of the movement features contained in such signature, a fine-grained perception of the Doppler effect is key. mmWave signals perfectly fit this requirement as they have higher carrier frequency, and consequently shorter wavelength, than sub-6 GHz systems, thus being much more sensitive to the frequency shifts and enabling higher velocity resolution. In [13] it is shown that the amplitude of the μD modulation is inversely proportional to the wavelength, so using mmWaves *amplifies* the capability of perceiving μD features with a RADAR device. This aspect is dealt with in more detail throughout the thesis and will be of fundamental importance in addressing the ISAC problem, in which we leverage communication waveforms for sensing, comparing the performance of mmWave and sub-6 GHz systems.

Despite the high appeal of μD effects, direct mathematical analysis of human movement-induced signatures is often infeasible, due to the complexity of the underlying motion and of the reflective properties of the human body. Indeed, a common approach consists in approximating the human body with a number of rigid objects with known geometry (e.g., cylinders or ellipsoids), modeling them as single scatterers moving according to the human gait pattern. Even in such a simplified approximation, which does not fully account for the complexity of the human body and its reflective properties, modeling the complex μD signature is highly non-trivial. For this reason, in many research works the tools of choice for μD feature extraction and processing are data-driven Machine Learning (ML) or Deep Learning (DL) algorithms. These have been shown to significantly outperform handcrafted feature extraction approaches in several tasks, e.g., activity recognition [3], [4]. Moreover, the powerful feature learning capabilities of deep Neural Networks (NNs) have enabled new applications that require even more fine-grained analysis of μD signatures, such as distinguishing person-specific movement patterns to map the RADAR reflected signal to a person's identity. In the next section, we introduce this aspect, which will be further investigated in Chapter 2 and Chapter 3.

Person identification from gait features using mmWave RADAR

Human gait has been classified as a *soft biometric* [15], meaning it is unique for each person. Differently from *hard biometrics*, however, such as fingerprints or DNA, it can not be used in high-stakes settings or to uniquely identify subjects among very large groups, e.g., more than $100 - 1,000$ people. Despite this, gait is difficult to fake, and it can be effectively analyzed even at distance and without requiring the subjects to collaborate. Several camera-based systems have been proposed to analyze human gait from videos, extracting features that embed the individual *way of walking* of a person and allow distinguishing her/him from others. mmWave RADAR-based gait recognition can be a good option to identify subjects in scenarios such as surveillance systems or individually-tailored smart home applications, where the number of people involved is in the order of a few tens, replacing or augmenting traditional camera systems.

Several characteristics make RADAR even more appealing than cameras in this sense. First, RADAR is immune to lighting conditions and weather, thus, differently from cameras, it can work in the dark or in the presence of smoke without any performance degradation. This is very important in security systems and search and rescue applications, where cameras often fail. Second, RADAR allows accurately reconstructing *distances* and μD features related to the movement velocity of each body part. On the contrary, distance measurements and 3D perception are not straightforward in vision systems, which only provide a bi-dimensional projection of the movement. Third, the use of radio waves opens interesting applications in contexts where the privacy of the users has to be preserved. The different nature of the information captured by RADAR sensing makes it less invasive than cameras, as only the movement-related information is retained.

Person identification from mmWave RADAR signatures is the connecting line of the works presented in the first part of this thesis. In Chapter 2, we propose and validate a multi-person tracking and identification system based on the integration between standard RADAR signal processing and DL. Our previous work [5] was the first one to perform simultaneous tracking and identification of multiple subjects concurrently moving in the same environment using μD signatures of gait. However, as is the case with many single-person approaches in the literature, e.g., [7] the overall system requires processing the full RADAR raw data cube to obtain highly detailed μD features. Our approach is different, as we explore the possibility of preprocessing the RADAR cube to extract a sparse *point-cloud* embedding low-resolution μD signatures together with 3D spatial information about the reflecting points. This makes the overall system lightweight and amenable to deployment on commercial edge computing devices. However, the reduced quality of the gait-related features makes the identification task much more challenging. The sparsity of radar point-cloud data can be a source of inaccuracy and prevents the direct use of standard DL convolutional architectures for the identification task, as point-clouds are *unordered* sets of points rather than structured inputs such as images. Therefore, the key challenges we solve are *(i)* the design of a suitable DL classifier that can process sparse RADAR point-clouds to provide accurate person identification, *(ii)* the integration of such classifier into the multi-person tracking system, to boost the tracking accuracy by utilizing the additional information on the person's identity, and *(iii)* solving challenges *(i)* and *(ii)* with a fast and lightweight system that can be implemented on commercial edge computers. Our solution is evaluated on a publicly available dataset of RADAR point clouds featuring 30 subjects, and on our own data including up to 8 subjects, obtained with a 77 GHz MIMO RADAR.

In Chapter 3, we leverage the system presented in Chapter 2 to jointly address the three tasks of body temperature screening, interpersonal distance monitoring, and contact tracing. This work stems from the timely need to accurately monitor indoor environments to counter contagious diseases, enabling the reconstruction of the chain of contacts in case of a contagion outbreak. Existing work has treated the three problems separately, often suffering from severe limitations in terms of usability and accuracy. We *jointly* address these challenges by *(i)* devising a novel method to fuse the information provided by an infrared thermal camera and a mmWave RADAR, and *(ii)* adapting our RADAR-based multi-person identification system from Chapter 2 to recognize people as they move across the rooms of an indoor space. The latter problem is much more complex than standard person identification, as people have to be recognized *on-the-fly* from only a few seconds of measurements, without having been previously observed by the system at training time. A similar task is known in the computer vision field as person re-identification (Re-Id) [16], [17]. We evaluate the proposed system on an extensive experimental campaign involving more than 20 subjects and joint infrared and mmWave measurements. An in-depth comparison to existing methods that separately perform interpersonal distance monitoring or temperature screening is provided, showing the superiority of our approach.

1.2 Integrated Sensing and Communication

RADAR and wireless communication systems have been progressing along independent, yet parallel tracks for several decades. However, the two share fundamental similarities, as any wireless communication device estimates the parameters of the surrounding propagation environment (*channel*) through probe RF signals, that are used to obtain the Channel Impulse Response (CIR). This is needed to mitigate the disruptive effects of so-called *multipath* reflections on buildings and objects. The channel estimation process can be considered as a *sensing* operation, as it allows perceiving some physical properties of the surroundings and inferring actionable insights about the context. This is the same underlying principle used in RADAR to localize and track targets of interest. However, RADAR systems are dedicated sensors, meaning they require ad-hoc costly deployment and the creation of supporting data processing infrastructure. Conversely, wireless networks count, as of 2021, more than 14 billion connected mobile devices already in place, continuously exchanging signals according to many standards, including the IEEE 802.11 (*Wi-Fi*) and 3GPP Fourth Generation (4G) and 5G New Radio (NR), across various licensed and unlicensed frequency bands [18]. In several scenarios (e.g., smart homes/buildings, offices, etc.), retrofitting standard communication devices with human and environment sensing capabilities is of great value to increase the scalability and ease of deployment of sensing systems. Thanks to their ubiquity, and their channel estimation capabilities, wireless communication systems hold the potential to become an unprecedentedly widespread, cheap, and pervasive sensing technology by applying the RADAR principle to communication signals. The ubiquitous deployment of such communication devices has sparked research interest towards developing ISAC technology, to avoid the cost of installing dedicated hardware while at the same time benefiting from communication capabilities.

Among the different standards and frequency bands, the most promising to be endowed with ISAC features are those working in the mmWave and the sub-Terahertz (sub-THz, $300-1000$ GHz) bands, among which are 3GPP 5G-NR, IEEE 802.11ad/, and the envisioned 3GPP 6G. This derives from the fact that RADAR can achieve higher resolution by transmitting signals having wider bandwidth. Therefore, the dominant trend of increasing wireless communication bandwidth to achieve higher data rates, by moving to higher and less crowded regions of the spectrum, could also enhance sensing capabilities. This has led to the identification of sensing as a key feature of next-generation 6G mobile networks and the creation of the IEEE 802.11bf standardization group [10], aimed at enabling sensing features in WLANs. While legacy Wi-Fi technology based on IEEE 802.11n and IEEE 802.11ac standards, working in the sub-6 GHz band, provides a viable means for environment and human sensing [19], and Human Activity Recognition (HAR) [20], [21], it suffers from intrinsic limitations due to its relatively low bandwidth. This prevents highly accurate distance measurements and multi-person localization and tracking in realistic scenarios. Conversely, by exploiting the available GHz-wide channels used in mmWave communication, ISAC can potentially sense objects' locations and movements with below centimeter-level accuracy, paving the way for countless applications in healthcare, security, navigation, autonomous driving, and many others [22].

In the second part of this thesis, we are concerned with the design of pervasive radio sensing systems that will extend the capabilities of upcoming Wi-Fi technology operating in the 60 GHz spectrum. Our target is to retrofit IEEE 802.11ay Physical Layer (PHY) to natively offer human and environment sensing services to end users. Most emerging systems, such as the ones presented in Chapters 2 and 3, are based on *dedicated* mmWave RADAR devices. These analyze the μD effect induced by human motion with high accuracy via specifically designed bursts of phase-coherent chirp signals [7], [23]. However, the extraction of μD signatures is difficult using standard communication devices and protocols, due to the lack of specifically designed waveforms and transmission modes. Extracting Doppler information from sequences of subsequent packets, as done in RADAR, is highly non-trivial due to the random and time-varying phase offsets between the transmitter and the receiver [24]. These offsets destroy the phase coherence across different packets, preventing the extraction of μD signatures which require a phase analysis across long sequences of subsequently transmitted signals. Moreover, communication traffic exhibits irregular and sparse patterns, which are not suitable for standard time-frequency analysis to extract μD signatures.

In Chapter 4, we propose the first way to effectively retrofit the IEEE 802.11ay standard to perform RADAR-like people tracking and μD signature extraction. Due to the high attenuation occurring at mmWaves, IEEE 802.11ay uses highly directional antennas for communication. For this reason, *beam-alignment* strategies have to be devised to find the best pair of beams to be used by the transmitter and the receiver to grant reliable communication. In IEEE 802.11ay, this is done via efficient *in-packet* beam training and tracking procedures [25], based on training (TRN) fields consisting of repetitions of complementary Golay sequences [26]. These fields are transmitted with different beam patterns, which allow determining which of the beam patterns is best for communication. Our main insight is that Golay sequences can be repurposed as RADAR pulses. Thanks to the 1.76 GHz channels used in IEEE 802.11ay, the resulting ranging accuracy allows reliable localization ad tracking of humans even in relatively crowded situations. In addition, the phase variations across reflections of subsequent packets can be used to extract the μD features of human movement. Note that these operations can be carried out *without modifying the underlying standard*. To do this, we propose to leverage the beam training process to track the position of each subject of interest, thanks to the possibility of switching beam patterns *within* the same communication packet. Conversely, the in-packet beam tracking is leveraged to reconstruct the μD spectrum. We implement the proposed method on a Field Programmable Gate Array (FPGA)-based SDR platform transmitting standard-compliant packets, and we address the tasks of multi-person tracking, HAR, and person identification. To enable comparison with widely studied sub-6 GHz sensing systems, a vast experimental campaign is conducted capturing RF data with our platform (mmWave) and a sub-6 GHz system based on IEEE 802.11ac routers. We show that our system significantly outperforms sub-6 GHz sensing and achieves performance comparable to a mmWave RADAR.

The main drawback of the framework presented in Chapter 4 is the need for continuous and dense sampling of the CIR. This requirement is imposed by the extraction of the human μD spectrum through conventional Short Time Fourier Transform (STFT), which requires uniform spacing between the samples of the analyzed signal. Our method shares this limitation with other approaches that perform target tracking or imaging [9], [27], [28]. The only practical way to make these systems coexist with communication is to alternate communication and sensing phases according to a time-division scheme, where regularly spaced, RADAR-like transmissions are performed during dedicated sensing periods. This is needed, in our system, to perceive the fine-grained μD effect of human motion, for which dense and regular sampling of the CIR is required, causing significant overhead and channel occupation. To solve this problem, in Chapter 5, we focus on enabling ISAC in realistic mmWave communication systems, by reusing existing communication traffic for sensing as much as possible and thus introducing only a minimal number of additional overhead and channel occupation. To this end, we propose the first mmWave ISAC system that reconstructs human μD signatures from *irregular and sparse* CIR samples obtained from realistic traffic patterns. The main idea behind this work stems from the observation that the high ranging accuracy of mmWaves, combined with the sparsity of the multipath environment at such high frequencies, causes the reflected signal to be sparse in the Doppler domain. Leveraging this fact, the number of packets that have to be collected to compute the μD spectrum can be significantly reduced by applying sparse reconstruction techniques such as Compressive Sensing (CS) [29], and the extraction can be made robust to irregular inter-packet time duration. This can be leveraged to exploit normal communication traffic for sensing purposes, reducing the need for injecting additional waveforms in the channel to the minimum. We evaluate this new, improved system using the same experimental setup used in Chapter 4, showing that the proposed sparse reconstruction technique can bring huge gains in terms of overhead and channel occupation reduction, while at the same time improving the quality of the resulting μD signatures.

1.3 Thesis outline

In the following chapters, we first delve into the analysis of mmWave RADAR-based human sensing algorithms design and validation, then we address the ISAC problem, leveraging a RADAR signal processing approach to solve communication-specific challenges.

In Chapter 2 we present the work in [6], that tackles computationally cheap simultaneous multi-person tracking and identification from sparse mmWave RADAR point-clouds. Chapter 3 instead refers to [30], where we addressed several challenges regarding the sensor fusion between mmWave RADAR and thermal cameras and the re-identification on-the-fly of unseen subjects using gait features from RADAR return signals.

Secondly, we turn to the problem of ISAC. Chapter 4 presents the first method to retrofit the IEEE 802.11ay beam training and tracking mechanisms to extract human μD signatures from communication packets, referring to [31]. In Chapter 5, we refer to [32], which extends the ISAC method in [31] to solve the problem of reusing sparse and irregular communication traffic for μD extraction, thus enabling much lower overhead and channel occupation for the sensing operations.

In Chapter 6, we draw some concluding remarks and propose future research directions stemming from the present work.

We conclude this introduction with a note on terminology. The acronym RADAR has become so widespread that often the word ”radar” is used for denoting a RADAR device or system. In this thesis, the two are used interchangeably.

2

Real-time People Tracking and Identification from Sparse mm-Wave Radar Point-clouds

2.1 Introduction

In this chapter, we begin our discussion on the use of mmWave radars for human sensing, focusing on the multi-person tracking and identification problem. Differently from existing solutions, our approach will be driven by practical considerations regarding computational complexity and real-time implementation. Our aim is to design and validate a real-time *multi-target* tracking and identification system running on constrained edge-computing devices[*] equipped with hardware accelerators (last generation Graphical Processing Units (GPUs)). Instead of working on the raw data obtained from the backscattered mmWave signal, as commonly done in the literature, we use *sparse point-clouds*. This makes it possible to implement our system on resource limited edge-computing devices. Point-clouds carry information about the three-dimensional spatial coordinates of the reflecting points, their velocity and the reflected power, and are obtained by employing detection algorithms at the radar processing unit, thus avoiding the need for transferring the *full raw data* from the radar to the edge computer. Due to their much lower data size, they bring advantages in terms of communication and computation at the connected processing device. Nonetheless, these advantages entail a *more challenging person identification task*: the sparsity of radar point-cloud data can be a source of inaccuracy and *standard DL architectures are inapt for learning from them*, as they rely on the reciprocal ordering of their input elements [33]. As a solution, we present a novel DL classifier, called Temporal Convolution Point-Cloud Network (TCPCN), which allows extracting meaningful order-invariant features from sparse point-cloud data.

[*] As an example, see the NVIDIA Jetson series.

The proposed system sequentially performs person tracking and identification, estimating the positions and the identities of humans as they freely move in an indoor space. For that, we use a low-cost Texas Instruments IWR1843BOOST mmWave, FMCW, MIMO radar and implement the required processing functions in real-time on a commercial edge-computing node (NVIDIA Jetson series). To carry out the person identification task, we combine standard tracking techniques, i.e., Kalman filter, with DL methods. This combined use of filtering and DL makes it possible to effectively capture the time evolution of the point-cloud representing each subject. Our main contributions are:

1. We build an end-to-end tracking and identification system that reliably operates in real-time at over 15 fps on a commercial edge-computing device paired with a low-cost mmWave radar. The approach reaches an accuracy of 91.62% in identifying up to three subjects (among a group of eight) freely and concurrently moving in a new indoor space, i.e., not seen at training time.

2. We propose a novel DL classifier, called TCPCN, that is tailored on mmWave radar point-cloud sequences and that is both accurate and fast. TCPCN contains a feature extraction block that obtains global information from the radar output at each time-frame and a block that exploits causal dilated convolutions [34] to recognize meaningful patterns in the temporal evolution of the features. Our model significantly outperforms state-of-the-art neural networks in this field in terms of classification accuracy and inference time.

3. The tracking phase of our system employs a Converted-Measurements Kalman Filter (CM-KF) that, in addition to estimating the position of the targets in Cartesian coordinates, also estimates the extension of the subject in the horizontal plane ($x - y$), considering him/her as an *extended object* rather than an ideal point-shaped reflector. This provides useful additional information that could be exploited by, e.g., occupancy or proximity based applications. In fact, knowing the extension of the subjects would be valuable for *(i)* smart-home applications that perform occupancy detection in certain areas, *(ii)* security systems in industrial settings, to estimate how close a person is to some dangerous area or machinery, *(iii)* detection systems (e.g., for automatic gates) that could quickly discern between cars, adults, kids or pets from their size. To the best of our knowledge, no earlier work uses Extended Object Tracking (EOT) within a point-cloud based tracking and identification system.

The novelty of the proposed solution stems from the following main points: the design and implementation of a novel DL-based neural network classifier working on time sequences of sparse point-cloud data, that is at the same time *highly accurate* and *fast*, the integration of tracking and identification phases, that in the literature on the subject are usually dealt with separately, the implementation and validation of the solution on a commercially available edge-computing platform with limited capacity.

The rest of the chapter is structured as follows. In Section 2.2, the literature on person identification using mmWave radars is reviewed, underlining the novel aspects in our approach. In

Section 2.3, the FMCW MIMO radar signal model is outlined, by also describing the procedure to extract the point-clouds. Our proposed framework is presented in Section 2.4. In Section 2.5, experimental results are shown, while concluding remarks are given in Section 2.6.

2.2 Related Work

In the last few years, person identification from backscattered mmWave radio signals has attracted a considerable and growing interest. Most of the research attention has been paid to processing human μD signatures as a means to distinguish among subjects, usually employing deep learning classifiers, applied to the μD spectrogram [7], [35]–[41]. Although this approach is robust and accurate, it presents some drawbacks. First, the extraction of μD signatures in case of multiple targets is a rather complex endeavor, and most of the above referenced solutions only work for a single-subject. In very few works, e.g., [41], the authors devised methods to single-out the contribution from multiple concurrent targets, obtaining the individual μD signatures. However, in the interest of obtaining highly accurate signatures, these previous algorithms dealt with *non-sparse* radar Range-Doppler-Azimuth (RDA) maps that require a large communication bandwidth to transfer the raw radio data from the radar to the processing device, preventing their implementation on low-cost embedded boards.

Only a few works so far have considered point-clouds obtained from a low-cost mmWave MIMO radar device. The sparsity of radar point-cloud data makes the identification task more challenging, as the specific features that identify each subject are more difficult to extract, and more sensitive to external disturbances. In [8], a recurrent neural network with Long Short-Term Memory (LSTM) cells is used for the identification. The overall accuracy obtained for 12 subjects is around 89%, and evidence that the system is able to distinguish between two concurrently walking subjects is provided. However, no evaluation of the accuracy is conducted when more than 2 subjects share the same physical space, nor by testing it in a different indoor environment (e.g., a new room) after its training. In addition, the point-cloud nature of the radar data is not fully exploited: the velocity and the received power are not used, and the classifier network requires the input data to be mapped onto a 3D voxel representation, which is inefficient and computationally expensive. The authors of [12] proposed a deep learning model that outperforms the bi-directional LSTM in [8] on their dataset. Two radar devices are used, transmitting and receiving simultaneously, leading to an increased field of view in case of blockage. However, robust methods are neither provided for tracking multiple subjects, e.g., Kalman or particle filtering [42], nor to reliably associate the detections (user identities) with trajectories. This seriously impacts the identification performance when multiple targets freely move in the monitored environment. In [12], it is in fact reported that the accuracy drops to 45% in a multi-target setting.

With the present chapter, we fill a literature gap, by designing a system that performs accurate tracking of multiple subjects from their point-clouds. Extended object tracking based on Kalman filtering is exploited in conjunction with a fast and novel domain-specific deep learning classifier. A tight integration of the tracking and identification modules is sought, towards enhancing the identification robustness and avoiding wrong identity associations and trajectory swaps. Moreover, and to the best of our knowledge, we are the first to provide an empirical study on the feasibility of operating the system in real-time on commercial edge-computing devices, and low-cost mmWave radars.

2.3 mmWave Radar Signal Processing

A Frequency-Modulated Continuous-Wave (FMCW) radar allows the joint estimation of the distance and the radial velocity of the target with respect to the radar device. This is achieved by transmitting sequences of linear *chirps*, i.e., sinusoidal waves with frequency that is linearly increased over time, and measuring the frequency shift of the reflected signal at the receiver. The frequency of the transmitted chirp signal is increased from a base value f_o to a maximum f_1 in T seconds. Defining the bandwidth of the chirp as $B = f_1 - f_o$, bandwidth B and chirp duration T are related through $\zeta = B/T$, and the instantaneous frequency of the transmitted signal is expressed as

$$f(t) = f_o + \frac{\zeta}{2}t, \quad 0 \leq t \leq T. \tag{2.1}$$

The phase of the transmitted signal is related to the instantaneous frequency by the following relation

$$\frac{1}{2\pi}\frac{d\varphi(t)}{dt} = f(t), \tag{2.2}$$

so it can be derived as

$$\varphi(t) = 2\pi\int_0^t f(t')dt' = 2\pi\left(f_o t + \frac{\zeta}{2}t^2\right). \tag{2.3}$$

Using Eq. (2.3), we can write the expression of the transmitted signal as

$$s(t) = \exp\left(j\varphi(t)\right) = \exp\left[j2\pi\left(f_o + \frac{\zeta}{2}t\right)t\right], \quad 0 \leq t \leq T. \tag{2.4}$$

The chirps are transmitted every T_{rep} seconds in sequences of L chirps each, so that the total duration of a transmitted (TX) sequence is LT_{rep}. A full sequence, termed *radar frame*, is repeated with period Δt. At the receiver, a mixer combines the received signal (RX) with the one transmitted, generating the IF signal, i.e., a sinusoid whose instantaneous frequency corresponds to the difference between those of the TX and RX signals. Each chirp is sampled with sampling period T_f (referred to as *fast time* sampling) obtaining M points, while L samples, one per chirp from adjacent chirps, are taken with period T_{rep} (*slow time* sampling).

The use of MIMO radar devices allows the additional estimation of the AoA of the reflections, by computing the phase shifts between the receiver antenna elements due to their different positions (i.e., their different distances from the target). This is referred to as *spatial* sampling, and enables the localization of the targets in the physical space. The radar device used in this chapter has $N_{\text{TX}} = 3$ transmitter and $N_{\text{RX}} = 4$ receiver antennas, that are equivalent to a virtual receiver array of $N_{\text{TX}}N_{\text{RX}} = 12$ antennas. The transmitting elements are arranged along two spatial dimensions, which we refer to as azimuth (AZ) and elevation (EL), and are used to transmit the chirp sequences according to a Time-Division Multiplexing (TDM) scheme. This enables the estimation of the EL and AZ angles of the reflecting points. In Section 2.3.1, we first consider one of the receiver elements, referring to it as *reference antenna*, and describe how the range and velocity of the subjects are estimated. In Section 2.3.2, we extend the discussion to multiple receiver antennas, showing how the AZ and EL AoAs are computed.

2.3.1 Range and Doppler information

Next, we show how to extract the range and velocity information from the received signal, focusing on the reference antenna. The signal reflected by a target is an attenuated version of the transmitted waveform with a delay τ that depends on the distance between the target and the radar and on their relative radial velocity.

Denoting by c the speed of light, and letting R and v respectively be the range and velocity of the target with respect to the radar device, the reflected signal delay is

$$\tau = \frac{2(R + vt)}{c}. \tag{2.5}$$

After mixing and sampling, the IF signal is expressed as [11]

$$y(m, l) = \alpha \exp\left[j\varphi_{\mathrm{IF}}(m, l)\right] + \mathrm{w}(m, l), \tag{2.6}$$

where m and l represent the sampling indices along the fast and slow time, respectively, α is a coefficient accounting for the attenuation effects due to the antenna gains, path loss and Radar Cross-Section (RCS) of the target and $\mathrm{w}(m, l)$ is a Gaussian noise term. The phase $\varphi_{\mathrm{IF}}(m, l)$ depends on the fast time and slow time sampling indices. By neglecting the terms giving a small contribution, an approximate expression for $\varphi_{\mathrm{IF}}(m, l)$ is written by introducing the quantities $f_d = 2f_o v/c$ and $f_b = 2\zeta R/c$, which respectively represent the Doppler frequency and the *beat* frequency of the reflected signal,

$$\varphi_{\mathrm{IF}}(m, l) \approx 2\pi \left[\frac{2f_o R}{c} + f_d l T_{\mathrm{rep}} + (f_d + f_b)\, m T_f\right]. \tag{2.7}$$

Samples of $y(m, l)$ can be arranged into an $M \times L$ matrix containing all the information provided by a single antenna for a given time frame. The frequency shifts of interest, which reveal the range and velocity of each reflector, can be extracted after applying a bi-dimensional Discrete Fourier Transform (DFT) along the fast time and slow time dimensions, followed by taking the square magnitude of each obtained complex value. The result of this process is often referred to as radar Range-Doppler (RD), and represents the received power distribution along the range of distances and velocities of interest.

The detection of the main reflecting points is performed using the Cell-Averaging Constant False Alarm Rate (CA-CFAR) algorithm on the range-Doppler maps [43], which consists in applying a dynamic threshold on each RD value (or *bin*), depending on the power of nearby *training* values. The use of an adaptive threshold introduces sparsity in the resulting set of detected points, as a point is retained (i.e., selected) only if its power is sufficiently larger than the average power of its neighbors.

In addition, a processing step is required to remove the reflections from static objects, i.e., the *clutter*. This operation is performed using a Moving Target Indication (MTI) high pass filter that removes the reflections with Doppler frequency values close to zero [43].

The detection and MTI processing steps return a sparse RD map containing N^{det} detected reflecting points: the position of each value along the fast time reveals the corresponding frequency in the IF signal $f_d + f_b \approx f_b$, while the peak along the slow time reveals the Doppler frequency f_d. For each detected point, the *observed* desired quantities are then expressed as follows (we indicate with the symbol Δ the corresponding resolution)

$$\tilde{R} = \frac{f_b c}{2\zeta}, \quad \Delta\tilde{R} = \frac{c}{2B}, \tag{2.8}$$

$$\tilde{v} = \frac{f_d c}{2f_o}, \quad \Delta\tilde{v} = \frac{c}{2f_o L T_{\mathrm{rep}} N_{\mathrm{TX}}}. \tag{2.9}$$

Additionally, from the RD map we obtain the reflected, received power from each detection, denoted by P^{RX}.

2.3.2 Azimuth and Elevation angles estimation

The complex-valued RD map of the radar illuminated range, before taking the square magnitude, is computed at all the receiving antenna elements, and presents a different phase shift at each antenna, due to its different distance from the target. This fact is referred to as *spatial diversity* of the receiver array, and can be exploited to estimate the azimuth and elevation angles of the targets.

Denote by d the distance between two subsequent antennas along the azimuth and elevation dimensions and by ψ_{AZ} and ψ_{EL} the corresponding experienced phase shifts, respectively. Moreover, let θ and ϕ be the AZ and EL angles of a reflecting point, while $\lambda = c/f_o$ is the base wavelength of the transmitted chirps. The following relations hold

$$\begin{aligned} \psi_{\mathrm{AZ}} &\approx \frac{2\pi}{\lambda} d \cos\phi \sin\theta, \\ \psi_{\mathrm{EL}} &\approx \frac{2\pi}{\lambda} d \sin\phi. \end{aligned} \tag{2.10}$$

To compute the phase shift values, two DFTs across the samples taken at the azimuth and elevation antennas in the virtual receiver array are computed, extracting the peak positions similarly to what described in Section 2.3.1 for beat and Doppler frequency. Finally, the Cartesian

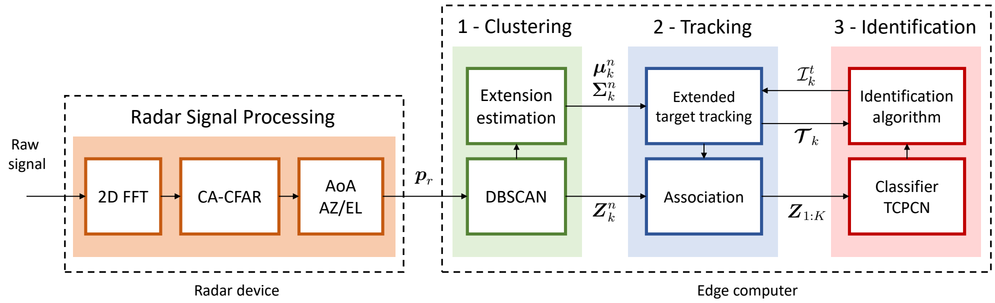


Figure 2.1: Block diagram of the proposed signal processing workflow: the raw radar data is processed on the radar device, extracting the sparse point-cloud representation of the environment, i.e., points $\mathbf{p}_r$, then (1) a clustering module groups the points $\mathbf{p}_r$ into the contributions from the different targets and estimates their position and extension, (2-3) tracking, data association and identification are *jointly* performed through an identification algorithm.

coordinates of each detected point are obtained using Eq. (2.10) as

$$\begin{aligned}
\tilde{x} &= \tilde{R}\cos\phi\sin\theta = \tilde{R}\frac{\lambda\psi_{\text{AZ}}}{2\pi d}, \\
\tilde{y} &= \sqrt{\tilde{R}^2 - \tilde{x}^2 - \tilde{z}^2}, \\
\tilde{z} &= \tilde{R}\sin\phi = \tilde{R}\frac{\lambda\psi_{\text{EL}}}{2\pi d}.
\end{aligned} \tag{2.11}$$

The vector describing a single detected reflecting point, $\mathbf{p}_r$, $r = 1, \dots, N^{\text{det}}$, has five components, containing the information on its Cartesian coordinates, its velocity and the reflected power: $\mathbf{p}_r = \left[\tilde{x}_r, \tilde{y}_r, \tilde{z}_r, \tilde{v}_r, P_r^{\text{RX}}\right]^T$.

2.4 System Design

The proposed system operates on discrete time steps, indicized by variable k, whose duration corresponds to the radar inter frame time Δt. At each frame, a set of N_k^{det} reflecting points $\mathbf{p}_r$ are obtained through the signal processing steps of Section 2.3. Our system sequentially performs the following operations on such points, see Fig. 2.1.

1. **Clustering and extension observation:** a density-based clustering algorithm is used to group the points detected by CA-CFAR into several clusters, each corresponding to a different subject present in the environment, see Section 2.4.1. The points associated with the different targets are then used to obtain *observations* of the subject's state, which according to our design includes his/her Cartesian position and *extension* in the horizontal plane $(x - y)$. The extension is modeled as an ellipse, that is determined by the spread (covariance) of the points in each cluster, Section 2.4.2.

2. **Tracking and data association:** a CM-KF [44] is used to estimate the position, velocity

and extension of the subjects in a Multi-Target Tracking (MTT) framework, processing the observations outputted by the previous step, Section 2.4.3. A set of trajectories, each corresponding to a human subject, are maintained and sequentially updated. The MTT association between new observations and trajectories is achieved using an approximation of the Nearest-Neighbors Joint Probabilistic Data Association (NN-JPDA) algorithm, see Section 2.4.4.

3. **Identification:** a deep NN classifier is applied to a temporal sequence of K subsequent point-clouds associated with each trajectory, with the objective of discerning among a set of Q pre-defined subject identities. The employed NN is called TCPCN, and is inspired by the popular PointNet architecture used for 3D point-cloud classification and segmentation [33]. TCPCN extends PointNet to the radar domain, by adding the velocity and received power information to the input and accounting for an additional block that handles the extraction of temporal features. Also, TCPCN is used in conjunction with an identification algorithm, which includes an exponential moving-average smoother and the Hungarian method, to jointly output a unique label for each trajectory: this combined use greatly improves the identification accuracy of the framework.

2.4.1 Point-cloud clustering – DBSCAN

Density-based clustering algorithms, as opposed to *distance*-based ones, group input samples according to their local density. One of the most widely used algorithms belonging to this category is Density-Based Spatial Clustering for Applications with Noise (DBSCAN) [45], which has been successfully applied to cluster radar point clouds in [8], [12], [41], [46]. The algorithm operates a sequential scanning of all the data points, expanding a cluster until a certain density connectivity condition is no longer met. The algorithm takes two input parameters, ε and m_{pts}, respectively representing a radius around each point and the minimum number of other points that must be inside such radius to meet the density condition. DBSCAN is only applied to the $x-y$ components of the detected points $\mathbf{p}_r$, namely, the Cartesian coordinates on the horizontal plane, as the different body parts of a subject can have very different velocity and reflected power values. We denote by $\{\mathbf{Z}_k^n\}_{n=1,\dots,D_k}$ the D_k clusters obtained at time step k by grouping the N_k^{det} detected points. In principle, there should be a distinct cluster for each human subject present in the environment, but due to several phenomena such as noise, imperfect clutter cancellation and blockage of the signal, a subject can go undetected even for several consecutive frames. DBSCAN was chosen for the following reasons: it is an unsupervised algorithm, i.e., the number of clusters (subjects) does not have to be known beforehand, it has a noise rejection quality that, together with its density-based clustering mechanism, allows a reliable and automatic separation of the reflections from distinct subjects, it has a low computational complexity, of about $\mathcal{O}\left(N_k^{\text{det}} \log N_k^{\text{det}}\right)$.

2.4.2 Subject Position and Extension Observations

Due to the high spatial resolution of mmWave radars, human subjects are detected as clusters containing tens of reflecting points. In the literature, the typical approach to their tracking has been to ignore the spatial extension of the targets, considering them as ideal point-shaped reflectors. In the present chapter, given a cluster of points $\mathbf{Z}_k^n$ selected by the DBSCAN clustering algorithm at time k, we instead obtain an estimate of the extension of the subject in the $x-y$ plane. As a first step, we define $\tilde{\mathbf{p}}_r = [\tilde{x}_r, \tilde{y}_r]^T$ and we normalize the received power values, P_r^{RX}, of the detected points in $[0,1]$. The spread of the points within each cluster around the cluster centroid provides a measure of the subject's extension. The centroid represents a noisy observation of the true position of the person, and is obtained as

$$\boldsymbol{\mu}_k^n = \sum_{r:\tilde{\mathbf{p}}_r \in \mathbf{Z}_k^n} P_r^{\text{RX}} \tilde{\mathbf{p}}_r, \tag{2.12}$$

where $\boldsymbol{\mu}_k^n = [\mu_{x,k}^n, \mu_{y,k}^n]^T$ and the received normalized powers P_r^{RX} act as weights. The covariance matrix, $\boldsymbol{\Sigma}_k^n$, contains information on the dimensions of the ellipse representing the extension of cluster n, and is obtained through the weighted sample covariance estimator,

$$\boldsymbol{\Sigma}_k^n = \sum_{r:\tilde{\mathbf{p}}_r \in \mathbf{Z}_k^n} P_r^{\text{RX}} \left(\tilde{\mathbf{p}}_r - \boldsymbol{\mu}_k^n\right) \left(\tilde{\mathbf{p}}_r - \boldsymbol{\mu}_k^n\right)^T . \tag{2.13}$$

The norms of the eigenvectors of matrix $\boldsymbol{\Sigma}_k^n$, denoted by $\tilde{\ell}_k^n$ and $\tilde{w}_k^n$ provide the axes lengths of the ellipse, while the orientation, $\tilde{\xi}_k^n$, has the same direction of the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}_k^n$.

2.4.3 Extended Object Tracking – Converted Measurements Kalman Filter

With the tracking step, we perform a sequential estimation of the *state* of the subjects present in the environment from their observed positions and extensions. To this end, we use a set of CM-KFs to establish a so-called *track* for each subject. A new Kalman Filter (KF) model is initialized for each detected cluster in the first frame received by the radar, while in successive frames, the tracks are maintained through the KF predict-update steps [42]. We denote by $\mathcal{T}_k^t$ the track with index t at time k, by $\mathcal{T}_k$ the set of currently maintained tracks, i.e., $\mathcal{T}_k = \{\mathcal{T}_k^t\}_{t=1,\dots,T_k}$, and by T_k its cardinality. We define the state of $\mathcal{T}_k^t$ as $\mathrm{x}_k^t = \left[x_k^t, y_k^t, \dot{x}_k^t, \dot{y}_k^t, \ell_k^t, w_k^t, \xi_k^t\right]^T$, which contains the true (and unknown) user's position (x_k^t and y_k^t), velocity ($\dot{x}_k^t$ and $\dot{y}_k^t$), extension (ℓ_k^t and w_k^t) and orientation angle (ξ_k^t). Each track is then defined as a tuple, $\mathcal{T}_k^t = \left(\hat{\mathrm{x}}_k^t, \mathbf{P}_k^t, \mathbf{Z}_{k-K+1:k}^t, \mathcal{I}_k^t\right)$, containing respectively the current state estimate, $\hat{\mathrm{x}}_k^t$, the associated error covariance matrix as computed by the KF, $\mathbf{P}_k^t$, the collection of the last K clusters associated with the track, $\mathbf{Z}_{k-K+1:k}^t$, to be fed to the NN classifier, and an integer $\mathcal{I}_k^t$ representing an estimate of the identity of the associated subject, at time k. The observation vector for a detected target n at time k is

$\mathrm{z}_k^n = \left[\mu_{x,k}^n, \mu_{y,k}^n, \tilde{\ell}_k^n, \tilde{w}_k^n, \tilde{\xi}_k^n\right]^T$.

The matching between any given cluster n and a corresponding track t ($n \leftrightarrow t$) is carried out using a specific procedure that will be detailed shortly in Section 2.4.4. For the sake of a concise notation, for the remainder of this section we drop the indices n and t, as the procedure that we describe next is carried out independently for each track (subject) once the matching $n \leftrightarrow t$ is performed.

Given the sequence of all collected measurements for a track up to time k, $\mathrm{z}_{1:k}$, the state estimation is carried out using the CM-KF. This approach assumes a posterior Gaussian distribution of the state given the sequence of measurements, i.e., $p(\mathrm{x}_k|\mathrm{z}_{1:k}) = \mathcal{N}(\hat{\mathrm{x}}_k, \mathbf{P}_k)$. To update $\hat{\mathrm{x}}_k$ and $\mathbf{P}_k$, a KF recursion [42] is applied using the measurements transformed in Cartesian coordinates from Section 2.4.2.

The model of motion that is used by the Kalman filtering block is defined by two matrices, $\mathbf{F}$ and $\mathbf{H}$. $\mathbf{F}$ is the transition matrix, connecting the system state at time k, x_k, to that at time $k-1$, x_{k-1}. $\mathbf{H}$ is the observation matrix, which relates the observation vector z_k to the true state x_k. Referring to $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and $\mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$ as the process noise and observation noise, respectively, a dynamic model of the system is

$$\mathrm{x}_k = \mathbf{F}\mathrm{x}_{k-1} + \mathbf{u}_k, \tag{2.14}$$

$$\mathrm{z}_k = \mathbf{H}\mathrm{x}_k + \mathbf{r}_k. \tag{2.15}$$

Denoting by blkdiag$[\mathbf{A}, \mathbf{B}]$ the block diagonal matrix with blocks given by matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\mathbf{F} = \text{blkdiag}\left[\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \otimes \mathbf{I}_2, \mathbf{I}_3\right], \tag{2.16}$$

and

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 3} \\ \mathbf{0}_{3\times 2} & \mathbf{0}_{3\times 2} & \mathbf{I}_3 \end{bmatrix}, \tag{2.17}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix, $\mathbf{0}_{n\times m}$ is an $n \times m$ all-zero matrix and $\otimes$ refers to the Kronecker product between matrices.

We assume the process noise $\mathbf{u}_k$ is due to a random acceleration a_k that follows a Gaussian distribution with 0 mean and variance σ_a^2, i.e., $a_k \sim \mathcal{N}(0, \sigma_a^2)$, leading to $\mathbf{u}_k = \mathbf{g}a_k$ with $\mathbf{g} = \left[\Delta t^2/2, \Delta t\right]^T$. The process noise covariance matrix is obtained as

$$\mathbf{Q} = \text{blkdiag}\left[\sigma_a^2 \mathbf{g}\mathbf{g}^T \otimes \mathbf{I}_2, \text{diag}\left(\sigma_\ell^2, \sigma_w^2, \sigma_\xi^2\right)\right], \tag{2.18}$$

with $\sigma_\ell^2, \sigma_w^2, \sigma_\xi^2$ being the constant process noise variances on the extension- and orientation-related coordinates of the state. The observation noise has covariance matrix given by

$$\mathbf{R}_k = \text{blkdiag}\left[\mathbf{R}'(\mathrm{x}_k), \text{diag}\left(\sigma_{\tilde{\ell}}^2, \sigma_{\tilde{w}}^2, \sigma_{\tilde{\xi}}^2\right)\right], \tag{2.19}$$

with $\sigma_{\tilde{\ell}}^2, \sigma_{\tilde{w}}^2, \sigma_{\tilde{\xi}}^2$ being the constant observation noise variances on the extension- and orientation-related coordinates of the state. For what concerns $\mathbf{R}'$, as radar measurements are obtained in polar coordinates, and then converted to the Cartesian space using Eq. (2.11), the measurement covariance matrix is time-varying as it depends on the current target's position. The sub-matrix $\mathbf{R}'$ accounts for the uncertainty in the Cartesian position observations, reflecting that an error on the AoA causes a higher uncertainty in Cartesian coordinates as the distance of the subject increases, due to the non-linear mapping between polar and Cartesian coordinates. In setting the uncertainty parameters for the measurements, we use a constant measurement covariance in polar coordinates, $\mathbf{R}_{\text{pol}} = \text{diag}\left(\sigma_R^2, \sigma_\theta^2\right)$, where R and θ are the distance and azimuth AoA, respectively introduced in Section 2.3.1 and Section 2.3.2. Hence, we use the transform $\mathbf{R}'(\mathrm{x}_k) = \mathbf{J}_{|\mathrm{x}_k}\mathbf{R}_{\text{pol}}\mathbf{J}^T_{|\mathrm{x}_k}$, where $\mathbf{J}_{|\mathrm{x}_k}$ is the Jacobian matrix of the conversion between polar and Cartesian coordinates, computed using the polar representation of the true subject state, x_k, which we approximate with $\mathrm{x}_k \approx \mathbf{H}\hat{\mathrm{x}}_{k-1}$. Although it can be seen that our conversion to Cartesian coordinates is biased, we remark that employing the unbiased conversion proposed in [47] did not lead to significant improvements. Note that, by the structure of the model matrices in Eq. (2.16) and Eq. (2.17), the kinematic part of the subject state and the extension part are entirely decoupled and do not interact during the CM-KF operations.

As a final remark about the KF model, with our approach the extension of the subject is explicitly accounted for as part of the state, fitting the point-clouds with ellipses, similarly to [48]. Although other approaches exist, such as using random matrices [49], [50], we found that our method leads to more accurate and meaningful extension estimates of the target's shape, due to the fast variability of radar point-clouds.

2.4.4 Data Association – NN-CJPDA

The association between new observations and tracks is needed *(i)* to correctly update the tracks with the observations generated by the corresponding subjects in a multi-target scenario, *(ii)* to correctly collect the sequence of the past K point-clouds associated with each subject, $\mathbf{Z}^t_{k-K+1:k}$.

To match tracks t to clusters n ($n \leftrightarrow t$), we use the NN-JPDA scheme. This method consists in computing the probability of each possible association between the D_k new clusters and the previous T_k tracks. These probabilities are then arranged into a $D_k \times T_{k-1}$ matrix of scores, $\mathbf{\Gamma}$, and the final assignment is done considering the association leading to the maximum overall probability, computed using the Hungarian algorithm [51]. The Hungarian algorithm uses the score matrix as input and solves the problem of pairing each track with only one cluster while maximizing the total score, entailing an overall complexity $\mathcal{O}((T_{k-1}D_k)^3)$.

To compute the probability of each match, i.e., the elements of matrix $\mathbf{\Gamma}$, we consider the widely adopted JPDA logic, using the approximate version of [52] called Cheap Joint Probabilistic Data Association (CJPDA). Exploiting the fact that the kinematic, extension and orientation parts of the state are decoupled in our framework, we apply CJPDA only using the kinematic state, as extension and orientation are more unreliable and could lead to association errors. Hence, in the following we refer to the kinematic part of the KF vectors and matrices only, i.e., to the components related to the Cartesian position and velocity of the targets.

The score matrix $\mathbf{\Gamma}$ is computed as follows (the time index k is omitted for a simpler notation). First, for all track-detection pairs the quantity G_{nt} is computed, which is proportional to the Gaussian function expressing the likelihood that observation n is produced by the subject corresponding to track t

$$G_{nt} = \frac{1}{\sqrt{\det \mathbf{S}_{nt}}} \exp\left[-\frac{1}{2}\boldsymbol{\nu}_{nt}^{T}\left(\mathbf{S}_{nt}\right)^{-1}\boldsymbol{\nu}_{nt}\right], \tag{2.20}$$

where $\boldsymbol{\nu}_{nt} = \hat{\mathrm{x}}^{t} - \mathbf{H}\mathrm{z}^{n}$ is the *innovation* brought by measurement z^{n} to the kinematic state of track t, $\hat{\mathrm{x}}^{t}$, and $\mathbf{S}_{nt} = \mathbf{H}\mathbf{P}^{t}\mathbf{H}^{T} + \mathbf{R}$ is its covariance matrix, obtained as part of the KF recursion. Second, the association probabilities for each track-detection pair are computed following [52], as

$$\Gamma_{nt} = \frac{G_{nt}}{\sum_{t=1}^{T_{k-1}} G_{nt} + \sum_{n=1}^{D_k} G_{nt} - G_{nt} + \beta}, \tag{2.21}$$

where the bias term β accounts for the possibility that no measurement is a good match for a specific track and is connected with the probability of missed detection. In this chapter, β is empirically set to $\beta = 0.01$, preventing the association of track-detection pairs with a low G_{nt} score.

2.4.5 Track management

The proposed system is robust to subjects that randomly appear on and disappear from the monitored space: these events may happen due to blockage of the radar signal at any point in time, or because the subject has moved in or out of the radio range. Blockage is a frequent problem in mmWave propagation and it happens frequently in multi-target scenarios, as users may block the radio signal with their own body. To deal with undetected subjects and new cluster detections which cannot be reliably associated with any existing track, while keeping the complexity of the system as low as possible, we follow a so-called `m/n` logic. In detail, a track is maintained if it received a match with any of the clusters detected by DBSCAN for at least `m` out of the last `n` frames. Similarly, cluster detections that are not associated with any existing track are initialized as new trajectories if they are detected for at least `m` out of the last `n` frames. In addition, to avoid tracks to merge when the subjects move too close to one another, the inter-track proximity is monitored. If the estimated Euclidean distance[†] between any two tracks $\mathcal{T}_k^t$ and $\mathcal{T}_k^{t'}$ becomes smaller than the DBSCAN radius parameter, ε, we remove the track having the largest determinant of the estimated error covariance, i.e., $\mathrm{argmax}_{j\in\{t,t'\}}(\det \mathbf{P}_k^j)$.

2.4.6 Point-cloud pre-processing

The point cloud sequence $\mathbf{Z}_{k-K+1:k}^{t}$ obtained from each CM-KF track is pre-processed before being sent to the NN classifier. The features of the points are standardized by subtracting their mean value and dividing by their empirical standard deviation. Moreover, the point-clouds must contain

[†] Obtained as $d(\mathcal{T}_k^t, \mathcal{T}_k^{t'}) = ((x_k^t - x_k^{t'})^2 + (y_k^t - y_k^{t'})^2)^{1/2}$.

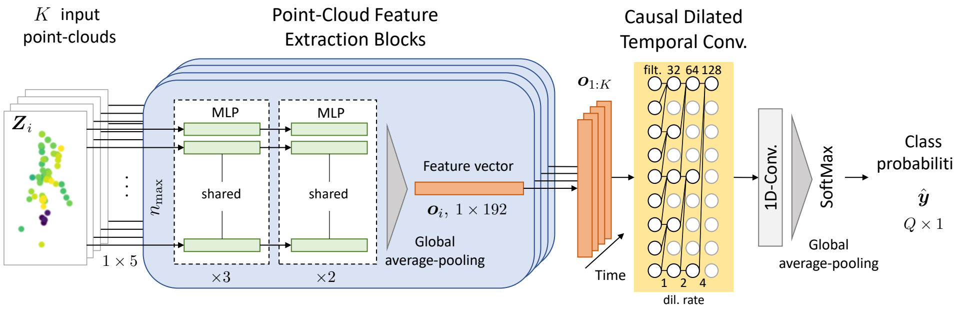


Figure 2.2: TCPCN – proposed DL-based classifier for subject identification: *(i)* a *point-cloud block* is applied to each individual time step to extract a feature vector, *(ii)* causal dilated convolutions are used to learn the temporal patterns in the sequence of feature vectors.

a fixed number of points before being sent to the TCPCN, as the latter is a feed forward neural network processing fixed size input vectors. We chose to limit the maximum number of points for a single time step to $n_{\max} = 100$. In case the number of points is greater than such maximum value, we randomly sample $n_{\max}$ points from the point-cloud without repetitions, in case there are fewer points than $n_{\max}$, some of the points are randomly repeated to reach the maximum value. The choice of $n_{\max}$ was made by analyzing the distribution of the number of detected points for different human subjects and empirically picking a suitable value: the selected $n_{\max}$ suffices to contain the point-clouds of all users in almost every frame in our experiments. Also, due to blockage and clutter, a subject may go undetected, especially in a multi-target scenario. If this occurs, the point-cloud data for the current frame is not collected for the blocked user and, in turn, is not sent to the NN classifier. A missed detection persisting over multiple radio frames may make the sequence of temporal features extracted for a subject by the NN less representative of his/her movement, and may ultimately degrade the identification performance of the algorithm. To ameliorate this, we propose an identification algorithm that jointly considers the outputs of the tracking block and of the classifier, as detailed in Section 2.4.9.

Considering that the TCPCN classifier is applied consistently to every track t at every time step k, in the following we simplify the notation denoting the pre-processed input point-cloud sequence $\mathbf{Z}^t_{k-K+1:k}$, of length K, by $\mathbf{Z}_{1:K}$.

2.4.7 Identification – Temporal Convolution Point-Cloud Network

The proposed classifier is designed to extract meaningful features from a temporal sequence of point-clouds, which is obtained as a result of the detection and tracking steps. The proposed architecture includes two processing blocks, termed Point-Cloud (PC) block and Temporal Convolution (TConv) block, and we refer to the full neural network as Temporal Convolution Point-Cloud Network (TCPCN), see Fig. 2.2.

Point-cloud Block

A number K of identical (same weights) feature extraction blocks is applied to the standardized input point-clouds, $\mathbf{Z}_i$, $i = 1, \dots, K$, of size $n_{\max} \times 5$, i.e., each composed of $n_{\max}$ reflecting points $\mathbf{p}_r$ (see Section 2.3.2). Each of such blocks implements a function $f_{\mathbf{W}}(\cdot)$, obtained as the cascade of a Multi-Layer Perceptron (MLP) [53] followed by a global average pooling operation, where $\mathbf{W}$ is a set of weights to be learned. Each reflecting point $\mathbf{p}_r$ in point-cloud $\mathbf{Z}_i$ (a vector of size 1×5), is fed to the first MLP layer and is independently processed from all the other $n_{\max}$ points in $\mathbf{Z}_i$, by one of $n_{\max}$ parallel branches. The MLPs located at the same depth share the same weights across all the points: there are 3 Fully-Connected (FC) layers with 96 units followed by 2 FC layers with 192 units. Each FC layer applies a linear transformation of the input followed by an Exponential-Linear Unit (ELU) activation function [54]. Batch normalization is used after each linear transformation [55] and right before the following non-linearity (ELU). The output feature vector from the last MLP layer from each branch has size 1×192. Global average pooling reduces this set of features to a single feature vector, $\mathbf{o}_i = f_{\mathbf{W}}(\mathbf{Z}_i)$, of size 1×192, by taking the average of each element across all the 100 parallel branches. The structure of function $f_{\mathbf{W}}(\cdot)$ is loosely inspired by the popular PointNet [33]. The key aspect of $f_{\mathbf{W}}(\cdot)$ is that it uses functions that are invariant to the ordering of the input points, by sharing the weights of the MLP and using suitable pooling operations. This ensures robustness and generality, because point-clouds that only differ in how the points are ordered will result in the same output. We underline that our TCPCN significantly differs from PointNet as the latter is designed to perform end-to-end classification and segmentation of *dense* 3D point clouds, whereas our $f_{\mathbf{W}}(\cdot)$ performs feature extraction from *sparse* 5D point-clouds.

Temporal Convolution Block

The sequence of feature vectors $\mathbf{o}_{1:K} = \{f_{\mathbf{W}}(\mathbf{Z}_i)\}_{i=1:K}$, each of dimension 192, is then fed to the PC block, which operates along the temporal dimension applying a function $h_{\mathbf{U}}(\cdot)$, where $\mathbf{U}$ is another set of weights. To extract temporal features efficiently, $h_{\mathbf{U}}(\cdot)$ contains temporal convolutions, which are a type of Convolutional Neural Network (CNN) layer [53] where the input is convolved with a uni-dimensional filter (or *kernel*) of learned weights in order to recognize temporal patterns. The output of the filters is organized into so-called *feature maps*, which become more and more complex and abstract with the depth of the layer. In TCPCN we use *causal dilated convolutions* [34], [56]. This technique consists *(i)* in masking the filters in such a way that neurons corresponding to a certain time step only depend on neurons corresponding to past time steps, i.e., they can not use future information, as done in [34], and *(ii)* in applying the convolution filters skipping blocks of $\delta - 1$ samples in the input, where δ is the so-called *dilation rate.* Formally, denoting a feature map as m and the filter as k, the output of a dilated convolution, $*_\delta$, between m and k is [56],

$$(\mathrm{m} *_\delta \mathrm{k})(s) = \sum_{i+\delta j=s} \mathrm{m}(i)\mathrm{k}(j). \tag{2.22}$$

The standard discrete convolution is obtained for $\delta = 1$. In the proposed TCPCN we employ 3 temporal convolution layers with filters of dimension 3 (also called *kernel* dimension) and dilation rates of $1, 2$ and 4, respectively. The applied filters are repeated along the feature vector components of the input, obtaining $32, 64$ and 128 feature maps at each layer, respectively.

The last layer of TCPCN is a temporal convolution layer that maps the extracted temporal features onto Q feature maps, each corresponding to one of the output classes. It applies a standard convolution with a kernel size of 3 and it is followed by a global average pooling to group the information from each feature map and obtain a single vector of dimension Q. Finally, a SoftMax function is applied, defined for a generic vector $\mathbf{x}$ as $\mathrm{SoftMax}(\mathbf{x})_i = e^{x_i}/\sum_j e^{x_j}$. The vector outputted by this last layer is denoted by $\hat{\mathbf{y}} = \mathrm{SoftMax}(h_{\mathbf{U}}(\mathbf{o}_{1:K})) = \mathrm{TCPCN}(\mathbf{Z}_{1:K})$ and its q-th elements represents the probability that the input point-cloud sequence belongs to class q.

2.4.8 Classifier Training and Inference

Loss Function

The loss function used is the categorical Cross-Entropy (CE), which is a standard choice in classification problems [53]. The CE compares the output of the last layer $\hat{\mathbf{y}}$ with the ground-truth identities of the subjects expressed in one-of-Q representation, $\mathbf{y}$: $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{q=1}^{Q} y_q \log(\hat{y}_q)$.

Training

To train TCPCN, we used the Adam optimizer with learning rate $\eta = 10^{-4}$ [53]. The process is stopped once the loss function computed on a validation set of data stops decreasing, a technique called *early stopping*. Overfitting is a severe problem in the context of radar point-clouds: the high randomness of the detected points and the sensitivity to different environments make the learning task challenging, especially when generalization to unseen environments is required. To reduce overfitting, several strategies were utilized: *dropout* [57] was applied to the output of the PC blocks, randomly dropping components of the feature vectors with probability $p_{\mathrm{drop}} = 0.5$, an L_2 regularization cost [53] on all network weights was considered, with parameter $\lambda_{L_2} = 10^{-4}$. The selection of the hyperparameters was carried out using a *greedy search* procedure.

Inference

During the inference (or prediction) step, TCPCN is used to obtain classification probabilities for each maintained track, $\mathcal{T}_k^t \in \mathcal{T}_k$, in the current time step k. We denote by $\tilde{\mathbf{y}}_k^t \in [0,1]^Q$ the vector that collects these probabilities. The prediction is carried out on a batch of T_k point-cloud sequences in parallel with a single pass of the data through the network, jointly obtaining $\{\tilde{\mathbf{y}}_k^t\}_{\mathcal{T}_k^t \in \mathcal{T}_k}$. Moreover, we apply weight quantization, [58], to 8 bit integer values to reduce even further the inference time and the memory cost of the model. It is worth noting that, due to the use of convolutions, TCPCN has a low number of parameters: in the PC block the weights are shared among the $n_{\max}$ parallel branches, while the PC block is a *fully convolutional* neural network, with no fully connected layers. Fully convolutional networks are typically very fast in terms of training and inference time compared to fully connected or recurrent neural networks and have fewer parameters (further analysis is carried out in Section 2.5.8).

2.4.9 Identification algorithm

After obtaining the output probabilities for each track from TCPCN, several problems still have to be tackled: *(i)* obtaining stable classifications, robust to the fact that subjects may turn or move in unpredicted ways which do not carry their typical movement signature, *(ii)* finding a method to compensate for the missing frames when subjects go undetected, which can cause classification errors, *(iii)* dealing with the uniqueness of the subject identities, as classifying the subjects independently and solely based on $\tilde{\mathbf{y}}_k^t$ may lead to assigning the same identity to multiple targets. To address these problems, we devised the procedure detailed in Alg. 2.1, which uses both the output of the tracking procedure and the classification probabilities provided by TCPCN to estimate the identities of the subjects in a stable and reliable way. The procedure acts as follows.

1. At the first time step $k = 1$, a vector $\mathbf{y}_1^t$ of size Q is initialized for each track $\mathcal{T}_1^t \in \mathcal{T}_1$, with all components equal to $1/Q$. $\mathbf{y}_1^t$ represents a stabilized vector of probabilities for each track.

2. At the generic time step $k > 1$, $\mathbf{y}_k^t$ is updated using Alg. 2.1, according to one of the two following rules:

   (a) if track $\mathcal{T}_k^t$ was detected in the most recent $K/2$ time-steps (line 1), TCPCN is applied to the corresponding sequence of point-clouds, obtaining the probability vector $\tilde{\mathbf{y}}_k^t$ (line 2). Hence, an exponentially weighted moving average procedure (line 6) is applied to mediate between the previous stable estimate $\mathbf{y}_{k-1}^t$ and the newly computed one $\tilde{\mathbf{y}}_k^t$, obtaining a new stable estimate $\mathbf{y}_k^t$ (normalized so that its elements sum to one, see line 7).

   (b) if track $\mathcal{T}_k^t$ was not detected in at least one of the most recent $K/2$ time steps (line 8), $\mathbf{y}_k^t$ is obtained as $\gamma\mathbf{y}_{k-1}^t$ with $\gamma < 1$ (line 9). In this way, we maintain the last reliable identification, but we progressively lower the confidence that we put on it over time. Note that after this step $\mathbf{y}_k^t$ does not longer resemble a probability distribution, as the sum of its elements is smaller than one.

3. To assign identities to subjects without repetitions, we build a matrix of scores $\mathbf{Y}_k$ with all vectors $\mathbf{y}_k^t$ belonging to each track (line 11). We compute the best assignment of the identities using the Hungarian algorithm on $\mathbf{Y}_k$, which guarantees that the joint maximum score is attained with a one-to-one mapping (line 13).

To avoid associating a label to a track if the corresponding probability is very low, in the identification process we use a slightly modified version of the Hungarian algorithm, which behaves as follows: first, we compute the associations using the standard Hungarian algorithm. Hence, if the probability of a certain association is below $p_{\text{conf}} = 0.1$, we set the identity of the considered track to *unknown*. In Alg. 2.1, this modified Hungarian algorithm is indicated as Hungarian$(\mathbf{Y}_k, p_{\text{conf}})$

Algorithm 2.1 Joint identification at time step k.

```
Input: Current set of tracks, $\mathcal{T}_k$, smoothing parameter, $\rho$, decay parameter, $\gamma$.
Output: Identities $\mathcal{I}_k^t$,   $\forall \mathcal{T}_k^t \in \mathcal{T}_k$.
Set $\mathcal{T}_k^{(s)} = \{\mathcal{T}_k^t \in \mathcal{T}_k$ s.t. $\mathcal{T}_k^t$ det. in the last $K/2$ frames$\}$
$\left\{\tilde{\mathbf{y}}_k^t\right\}_{\mathcal{T}_k^t \in \mathcal{T}_k^{(s)}} \leftarrow$ TCPCN $\left(\left\{\mathbf{Z}_{k-K+1:k}^t\right\}_{\mathcal{T}_k^t \in \mathcal{T}_k^{(s)}}\right)$
Initialize $\mathbf{Y}_k = \mathbf{0}_{T_k \times Q}$
for $\mathcal{T}_k^t \in \mathcal{T}_k$
    if $\mathcal{T}_k^t \in \mathcal{T}_k^{(s)}$
        $\mathbf{y}_k^t \leftarrow (1-\rho)\,\tilde{\mathbf{y}}_k^t + \rho \mathbf{y}_{k-1}^t$
        normalize $\mathbf{y}_k^t$
    else
        $\mathbf{y}_k^t \leftarrow \gamma \mathbf{y}_{k-1}^t$
    end if
    $(\mathbf{Y}_k)_{t,:} \leftarrow \mathbf{y}_k^t$
end for
$\mathcal{I}_k^t \leftarrow$ Hungarian$(\mathbf{Y}_k, p_{\text{conf}})$, $\forall\, \mathcal{T}_k^t \in \mathcal{T}_k$
```

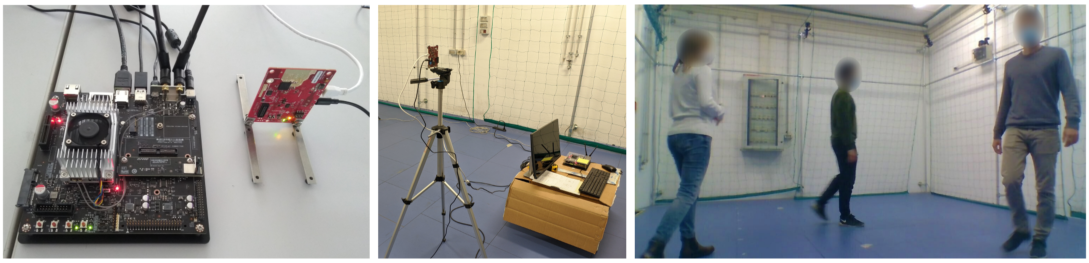
Algorithm 2.2 Tracking error correction at time step k.

```
Input: Current set of tracks $\mathcal{T}_k$.
Output: Updated set of tracks $\mathcal{T}_k'$.
1: for $\mathcal{T}_k^t \in \mathcal{T}_k$
2:     if $\mathcal{I}_k^t \neq \mathcal{I}_{k-1}^t$
3:         initialize new track $\mathcal{T}_k^j$ using $\mathrm{x}_k^t, \mathbf{P}_k^t$ and $\mathcal{I}_k^t$
4:         $\mathcal{T}_k' \leftarrow \{\mathcal{T}_k \setminus \mathcal{T}_k^t\} \cup \{\mathcal{T}_k^j\}$
5:     end if
6: end for
```

to highlight that the result is a function of the score matrix $\mathbf{Y}_k$ and of the confidence threshold p_{conf}.

With Alg. 2.1, we jointly exploit the information from the classifier (vector $\tilde{\mathbf{y}}_k^t$) and the tracking step ($\mathcal{T}_k^{(s)}$) to improve the identification performance.

Alg. 2.2 deals with errors in the tracking procedure, using the identity information available for each track. Tracking or association errors may happen during a blockage event involving two subjects (*blocker* and *blocked* in the following): for example a blocked subject may be erroneously associated when he/she becomes detectable again while being close to the blocker. These errors are dynamically corrected by analyzing the output of Alg. 2.1. Specifically, when the identity of a track $\mathcal{T}_k^t$ changes, we assume that this is an indication of a tracking error of the above-mentioned type (see line 2 of Alg. 2.2). In this case, this track is removed from the set of tracks that are maintained (line 4). At the same time, a new track $\mathcal{T}_k^j$ is initialized using the new identity $\mathcal{I}_k^t$, a new track index j (not yet used) and the current variables (state and covariance) associated with the old track $\mathcal{T}_k^t$ at time k (line 3). The new track $\mathcal{T}_k^j$ is then added to the set of maintained tracks (line 4). Note that, the memory $\mathbf{Z}_{k-K+1:k}^t$ (past frames) is not attached to the new track, which is started anew.



(a) Jetson board (left) and radar (right).

(b) The mounted setup in the test room.

(c) Three subjects walking in the test room.

Figure 2.3: Overview of the experimental setup.

2.5 Experimental results

In this section, we present results obtained by evaluating our tracking and identification method on

1. the mmGait dataset described in [12], available at `https://github.com/mmGait/people-gait` (Section 2.5.1).
2. Our own dataset, featuring 8 subjects (Section 2.5.2). This dataset was collected from our own measurements, implementing the proposed system on an NVIDIA Jetson TX2[‡] board paired with a Texas Instruments IWR1843BOOST mmWave radar[§] operating in the 77 – 81 GHz band.

The Jetson board mounts an NVIDIA Tegra X2 GPU accelerator, the radar device is connected to it via USB and the communication is performed via Universal Asynchronous Receiver-Transmitter (UART) ports, as shown in Fig. 2.3a. A camera was used to collect a video of the scene during the measurements and to label the dataset with the correct identities of the subjects. This setup poses some severe limitations on the amount of data that can be transferred in real-time to the NVIDIA processing device. Note that a more advanced solution such as an Ethernet connection would require additional hardware at an extra cost[¶]. The full system has been implemented in Python, using the TensorFlow library for the neural network classifiers. In Tab. 2.2, the system parameters used in the evaluation are summarized.

2.5.1 Evaluation on the mmGait dataset

To assess the capabilities of TCPCN to effectively extract human gait features from point-cloud sequences, we test it on the publicly available mmGait dataset [12], which contains measurements from two different evaluation rooms, `room_1` and `room_2`, including respectively 23 and 31 different

[‡] `https://developer.nvidia.com/embedded/jetson-tx2`

[§] `https://www.ti.com/tool/IWR1843BOOST`

[¶] `https://www.ti.com/tool/DCA1000EVM`

Setup		TCPCN (ours)		mmGaitNet [12]	
Room	# subj.	linear	free	linear	free
room_1	10	**92.07**	**70.31**	90.0	45.0
room_1	15	86.21	68.36	–	–
room_1	20	**83.37**	63.97	80.0	–
room_2	30/29	89.34	64.73	–	–

Table 2.1: Evaluation results on the mmGait dataset [12]. We report the accuracy (%) obtained by mmGaitNet according to the original paper [12] and the accuracy of our TCPCN, highlighting the best performance with a **bold** font. In the table, two columns show the results for linear and unconstrained motion. The dataset contains 29 subjects for room_2 in the free motion case, and 30 in the linear motion case. The symbol "–" is used for those cases for which no accuracy value is provided in [12].

subjects. The dataset contains sequences where subjects are constrained to walk along straight lines in front of the radar, and other sequences where they walk freely.

Next, we present a comparison between our neural network classifier, TCPCN, and the CNN proposed by the authors of mmGait, denoted by mmGaitNet [12]. The accuracy results obtained by TCPCN on a superset of the tests conducted by the authors in [12] are shown in Tab. 2.1. We stress that, for the sake of a fair comparison, for these results we just compared TCPCN with the CNN of [12], without using our algorithms Alg. 2.1 and Alg. 2.2, as they would provide an additional performance increase.

For the results in Tab. 2.1, we consider the mmGait traces recorded by a single TI IWR6843[‖] radar working in the 60 – 64 GHz frequency band. The measurements for each subject are split according to a 80% – 20% proportion to obtain training and test sets, as done in [12].

TCPCN outperforms mmGaitNet in all the considered cases. The gap is particularly large in case the subjects walk freely: in this case, mmGaitNet reaches an accuracy of 45% on 10 subjects, as compared to an accuracy of 70.31% for TCPCN. This difference is due to the high variety of patterns that occur in the presence of unconstrained motion. TCPCN is more robust to such variability thanks to its invariance to the ordering of the points in the data cloud. The obtained performance on 30 subjects is encouraging, leading to identification accuracies as high as 89.34% and 64.73% for linear and unconstrained motion, respectively. This shows that gait-based identification systems employing mmWave radar sensors hold the potential of scaling to scenarios where the number of users is in the order of a few tens. Finally, we point out that the accuracy with 30 subjects being higher than that with 15 and 20 is probably due to the fact that room_2 contains subjects who are more easily distinguishable than those from room_1.

2.5.2 Proposed dataset description

To further validate the proposed system, we built our own dataset using four different rooms: three to collect training data and one for testing purposes. This arrangement of data and rooms

[‖] https://www.ti.com/tool/IWR6843ISK

System parameters		
Antenna el. spacing	d	1.948 mm
Number of TX antennas	N_{TX}	3
Number of RX antennas	N_{RX}	4
Start frequency	f_o	77 GHz
Chirp bandwidth	B	3.072 GHz
Chirp duration	T	60 μs
Chirp repetition time	T_{rep}	68 μs
No. samples per chirp	M	256
No. chirps per seq.	L	64
Frame rate	$1/\Delta t$	14.92 fps
ADC sampling frequency	$1/T_f$	5 MHz
Range resolution	$\Delta\tilde{R}$	4.88 cm
Velocity resolution	$\Delta\tilde{v}$	14.9 cm/s
DBSCAN radius	ε	0.4 m
DBSCAN min. cluster dim.	m_{pts}	10
Meas. range std	σ_R	0.03 m
Meas. az. angle std	σ_θ	$\pi/24$ rad
Meas. ext. std	$\sigma_{\tilde{\ell}}, \sigma_{\tilde{w}}$	0.05 m
Meas. orient. std	$\sigma_{\tilde{\xi}}$	$\pi/6$ m
Process noise std	σ_a	8 m/s^2
Process ext. std	σ_ℓ, σ_w	0.001 m
Process orient. std	σ_ξ	$\pi/24$ m
CJPDA bias term	β	0.01
`m/n` logic parameters	`m/n`	10/30
Max point-cloud dim.	$n_{\max}$	100
Input time-steps	K	30
Moving avg. parameter	ρ	0.99
Decay parameter	γ	0.999
Dropout probability	p_{drop}	0.5
Regularization parameter	λ_{L_2}	10^{-4}
Learning rate	η	10^{-4}

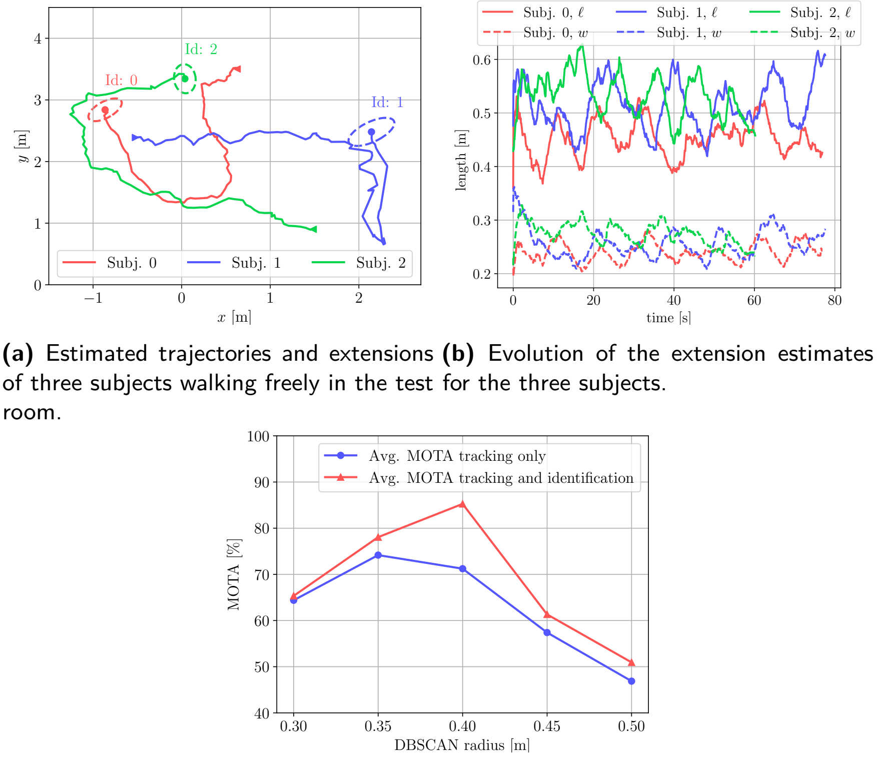
Table 2.2: Summary of the parameters of the proposed system.

was intentionally adopted to asses the generalization capabilities of the proposed system. Eight subjects were involved in the measurements, see Tab. 2.3.

Training: the training rooms are two research laboratories, of size 8×8 meters and 8×3 meters, respectively, containing desks, furniture and technical equipment, and a furnished living room of size 8×5 meters. In the first room, due to space limitations, the area used for the training measurements is a rectangular space of size 3×5 meters. To collect the training data, one subject at a time walked freely for an amount of time ranging from 1 to 5 minutes. Note that, in all our measurements the subjects are allowed to cover a distance of up to 6 m from the radar, within its

Subject	Age	Height [m]	Sex	ℓ [cm]	w [cm]	Frames
0	26	1.63	F	43	22	33,339
1	26	1.76	M	52	23	33,514
2	25	1.85	M	52	24	36,126
3	26	1.72	M	46	16	18,668
4	28	1.69	M	45	22	19,035
5	25	1.61	F	43	20	27,674
6	63	1.77	M	50	24	22,039
7	63	1.58	F	41	20	16,925

Table 2.3: Details on the subjects involved in the measurements.



(a) Estimated trajectories and extensions of three subjects walking freely in the test room.

(b) Evolution of the extension estimates for the three subjects.

(c) Average Multiple Object Tracking Accuracy (MOTA) using tracking only or joint tracking and identification.

Figure 2.4: Tracking system evaluation.

field-of-view of ±60°.

The measurement campaign was repeated across different days, acquiring from 20 to 40 minutes of data per subject. Taking into account different days, we aimed at reducing the effect of clothing or daily patterns in the way of walking. Prior to the actual training phase, the point-clouds data were pre-processed as described in Section 2.4.6 and grouped into sequences of $K = 30$ consecutive frames, leaving an overlap of 20 frames between different sequences. To reduce overfitting, we artificially augmented the training data by applying random shuffling of the points in each point-cloud and adding random noise to each point, drawn from a uniform distribution in the interval $[-0.1, 0.1]$. To select the neural network hyperparameters, a portion of the training data (one sequence of approximately $2,250$ frames per target) was used as a validation set.

Test: the test room is a 7 × 4 meters research laboratory, whose measurement area is free of furniture (see Fig. 2.3). We stress that, while training is performed on up to 8 *single subjects*, all our test sequences include multiple targets concurrently moving in the test environment. This leads to blockage events, i.e., when a subject occludes the line-of-sight (LoS) between the radar and another target, resulting in bursts of frames where the blocked subject goes undetected.

The measurement sequences contained in the test dataset are split as follows:

1. 10 sequences of 80 seconds ($1,200$ frames) with 3 subjects. These are further split into 5 sequences where the subjects were constrained to walk following a linear movement at their preferred speed (back and forth across predefined linear paths), and 5 sequences where they could walk freely, following any trajectory in the available space, as shown in Fig. 2.3c. This leads to unpredictable trajectories that can cover the whole field-of-view of the radar sensor ($\pm 60^\circ$) and distances up to 6 m. Moreover, in all our experiments user trajectories intersect frequently, leading to ambiguities in the data association, and making tracking more challenging.

2. 10 sequences of 80 seconds with 2 subjects, split into 5 sequences with a linear walking movement, and 5 sequences where the subjects walk freely.

2.5.3 Tracking phase evaluation

In Fig. 2.4a, we show example trajectories followed by the three targets in one of the test sequences. In this experiment, the CM-KF succeded in identifying and reconstructing the trajectory of each target, even in the presence of complex and strongly non-linear movement. The NN-CJPDA data association logic was found to be very robust, as long as the targets are correctly separated by DBSCAN into disjoint point-cloud clusters.

In all the test measurements the main difficulty faced by the system was that of handling blockage events that span over a large number of frames, e.g., more than $2-3$ second long. The number of such events increases significantly when more subjects are added to the monitored environment. We empirically assessed that, using a single radar sensor with the resolution and communication capabilities considered in this chapter, going beyond three freely moving subjects at a time in such a small indoor environment leads to insufficient tracking and identification accuracy due to blockage. This is coherent with the findings in the literature, e.g. [12], where two radars placed in different locations were used to compensate for these facts.

Fig. 2.4b shows the results of the extension estimation across a full test sequence for all subjects. The expected shape enclosing a human target is correctly estimated: the ellipse axes are coherent

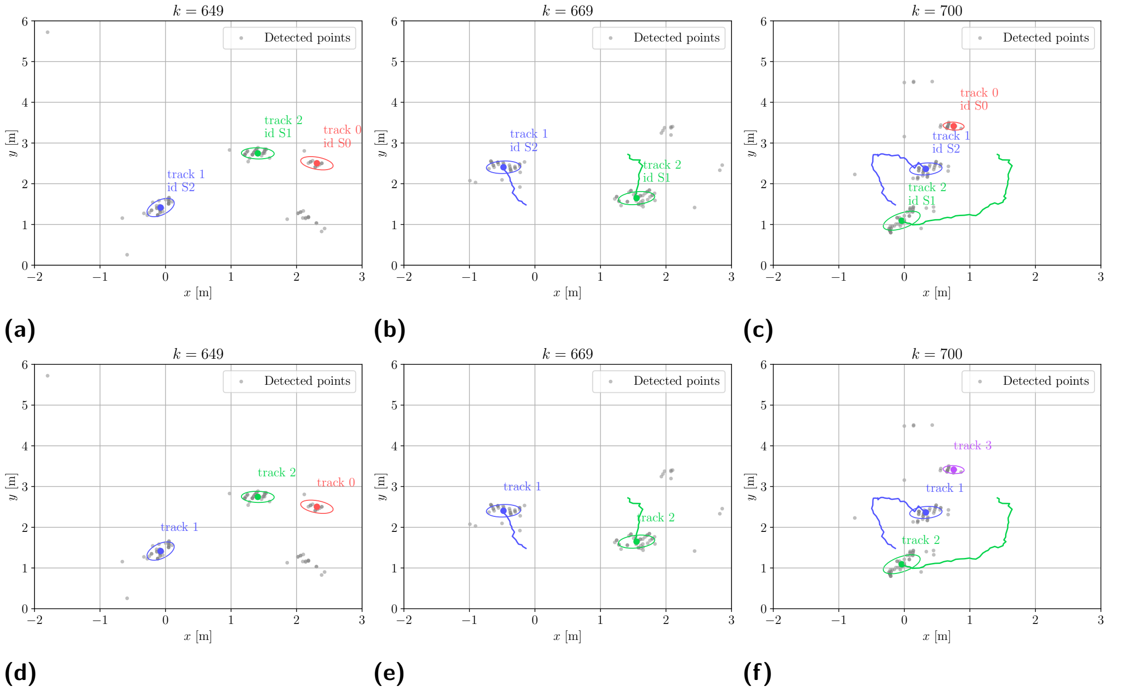


Figure 2.5: Proposed identification algorithm (a - b - c) compared to a standalone tracking approach (d - e - f) on the $x - y$ plane. Subject 0 (S0) is lost at time $k = 669$ and tracked again at time $k = 700$. By joint use of tracking and identification algorithms, the new track 3 is correctly re-associated with S0 (c), i.e., track 3 is mapped back onto track 0. Instead, the sole use of tracking would lead to the initialization of a new track for the same subject (f), causing a mismatch.

with typical shoulder widths, ℓ, and thorax widths along the sagittal plane, w. The estimated value varies depending on the position of the target with respect to the radar: this is due to the fact that the received point-clouds contain a smaller number of points as the distance increases, due to propagation losses. Despite this fact, the average values are still proportional to the true subjects' extensions, as it can be checked by comparing Fig. 2.4b with Tab. 2.3.

To evaluate the capability of the proposed system towards tracking human subjects and the improvement brought by combining tracking and identification algorithms, we use the popular MOTA metric [59]. The MOTA conveniently summarizes the ratio of missed targets (miss), false positives (fp) and track mismatches (mm), over the number of ground truth targets (gt) in each time frame k of the test sequence, formally,

$$\mathrm{MOTA} = 1 - \frac{\sum_k (\mathrm{miss}_k + \mathrm{fp}_k + \mathrm{mm}_k)}{\sum_k \mathrm{gt}_k}. \tag{2.23}$$

The value of gt_k was obtained from a reference video, as mentioned at the beginning of Section 2.5.

In Fig. 2.4c, we show the MOTA obtained for different values of the DBSCAN radius, ε, for the NN-JPDA algorithm and our method, where NN-JPDA is used in conjunction with Alg. 2.1

	Test 2 sub. train 3			Test 3 sub. train 3			Test 3 sub. train 8		
	linear	free		linear	free		linear	free	
[%]	Id. acc.	Id. acc.	MOTA	Id. acc.	Id. acc.	MOTA	Id. acc.	Id. acc.	MOTA
Seq. 1	98.67	99.61	98.71	100	99.67	99.06	96.95	92.35	99.06
Seq. 2	100	99.75	98.71	100	99.91	84.14	99.81	96.17	84.14
Seq. 3	95.26	96.91	86.42	100	91.79	94.11	100	88.14	90.19
Seq. 4	99.54	100	99.62	90.43	100	76.36	90.43	89.46	76.36
Seq. 5	99.34	100	97.96	99.37	92.44	72.61	97.04	92.02	72.61
Average	**98.56**	**98.98**	**96.28**	**97.96**	**96.76**	**85.26**	**96.85**	**91.62**	**84.47**

Table 2.4: Accuracy and MOTA obtained with 2 and 3 subjects moving in the test room. We report the results both when the subjects follow linear trajectories ("linear") and when they move freely ("free"). With "Test x sub. train y" we denote the fact that the TCPCN used for the identification was trained on the single-target measurements of y subjects and tested on multi-target sequences containing x subjects simultaneously.

and Alg. 2.2 (subject identification and label correction). Note that, with the standard NN-JPDA tracking algorithm, when a track is deleted and re-initialized, it is counted as a mismatch in Eq. (2.23), significantly lowering the MOTA. Moreover, data association errors can lead to track swaps when the trajectories of two subjects intersect. The MOTA obtained in this case is plotted as a blue curve in Fig. 2.4c. The red curve instead represents the improved MOTA, obtained by *(i)* merging together all the tracks associated with the same subject's identity, as described in Section 2.4.9, and *(ii)* correcting track swaps using Alg. 2.2. For the sake of clarity, in Fig. 2.5 we exemplify step *(i)*, which significantly improves the results by mitigating the effect of losing and re-initializing tracks.

From Fig. 2.4c, we see that for the optimal value $\varepsilon = 0.4$ m, the integration of tracking and identification provides an improvement of almost 20% in terms of MOTA. Remarkably, this is obtained at almost no additional complexity, by just feeding back the identity information to the tracking block.

2.5.4 Accuracy results

In Tab. 2.4, we report the person identification accuracy obtained with the proposed method on the test sequences described in Section 2.5.2. For the unconstrained walks, we also report the corresponding MOTA. The per-subject identification accuracy is computed using the time-steps in which the subject is correctly tracked, and is defined as the fraction of time-steps where a subject, besides being tracked, is also correctly identified. The final accuracy on a test sequence is obtained by taking the average accuracy on each subject, weighted by the total number of frames in which he/she is detected and tracked by the system.

In our tests, the number of subjects used for training is set as either 3 or 8 to assess how the system performs with an increasing number of targets. In Tab. 2.4, this is indicated with "Test x sub. train y", where x and y respectively refer to the number of subjects in the training set and those who are simultaneously present in the test data. The accuracy ranges from a maximum of

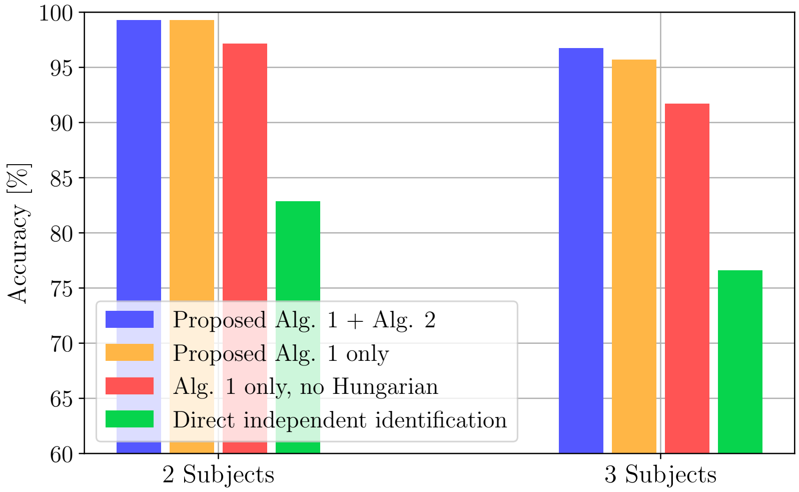


Figure 2.6: Accuracy of the proposed identification algorithm.

98.98% down to 91.62%, with the latter achieved for the most challenging case where 3 concurrent subjects have to be identified among a set of 8.

Differently from the results on mmGait (see Section 2.5.1), there are no significant deviations in the identification performance between linear and unconstrained motion. This is due to the proposed identification algorithm, which lowers the effect of turns and non-linear movements that are likely to impact the classification accuracy. The MOTA is instead significantly lower with three targets, because of the more frequent blockage events (more misses and mismatches).

Fig. 2.6 shows the average accuracy obtained over the free-walking test sequences by *(i)* using the proposed solution (Alg. 2.1 and Alg. 2.2), *(ii)* using Alg. 2.1 only, *(iii)* using Alg. 2.1 without the Hungarian method, and *(iv)* identifying each subject at each time step k by solely using the point-cloud data at time k, and estimating the identity as $\operatorname{argmax} \tilde{\mathbf{y}}_k^t$. For this evaluation the TCPCN was trained on 3 subjects. Note that with 2 subjects *(i)* and *(ii)* lead to about the same performance, but Alg. 2.2 leads to a slight improvement with 3 subjects, as tracking errors caused by track swaps due to blockage are more frequent in this case.

2.5.5 Impact of temporal filtering parameters

Now, we analyze the impact of K and ρ, i.e., the number of input time steps and the moving average smoothing parameter, respectively. These parameters are intimately connected, as they both control the dependence of the current output on past frames. In Fig. 2.7, we show the average accuracy computed on 10 different trainings of TCPCN with $Q = 3$ subjects, when tested on 3 subjects moving freely. The shaded areas represent 95% confidence intervals. In the abscissa, we vary K, plotting a different curve for several selected values of ρ. Lower values of ρ, e.g., 0.8 or 0.9, lead to a lower performance, as the memory of the moving average filter in these cases is too short to introduce stability in the classification (it corresponds to 5 and 10 time steps for $\rho = 0.8$ and

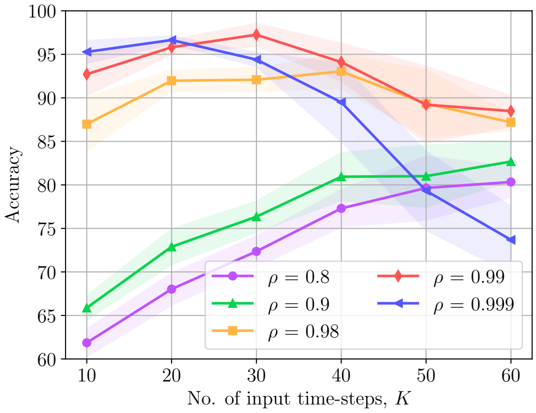


Figure 2.7: Effect of varying K and ρ on the identification accuracy.

0.9, respectively) and high values of K are required to get an accuracy beyond 80%. Increasing ρ has the effect of moving the point of maximum accuracy towards lower values of K. From our results, we recommend using $K = 30$ (two seconds of radar readings) and $\rho = 0.99$, as these values lead to the best average accuracy while keeping the system sufficiently reactive, with a moving average memory of approximately 100 time steps (between 6 and 7 seconds).

2.5.6 Importance of point-cloud features

In Tab. 2.5 we show the accuracy results of the sole TCPCN (no Alg. 2.1 and Alg. 2.2) considering 8 single targets, by leaving out some of the point-cloud features in $\mathbf{p}_r$. Specifically, we trained and tested the NN by selectively leaving out the received power (no-P), the velocity (no-v), the z coordinate (no-z) or the $x - y$ coordinates (no-xy). This evaluation provides insights on the importance of each of these features towards identifying the subjects. In particular, removing the velocity, $x - y$ or z coordinates led to the largest reduction in accuracy, suggesting that these carry the most useful information. In addition, Tab. 2.5 proves that our method mostly relies on movement-related features rather than on the reflectivity of the target (related to the received power). We remark that this is key to gain robustness to reflectivity changes due to different clothing or other environmental factors, and the lower importance of certain features is enforced by the learning procedure, which has automatically learned it by processing data from the same subjects across different days (wearing different clothes, etc.) and environments.

2.5.7 Real-time implementation requirements

Operating the proposed system in real-time poses constraints on the execution time of each processing block, and on the choice of the size and structure of the NN classifier. We measured the computation time needed by each block, respectively denoting by t_p the time needed to run the point-cloud extraction module running on the radar device (including the chirp sequence transmission, three DFTs along the fast time, slow time and angular dimension and the CA-CFAR

	all	no$-\mathbf{P}$	no$-\mathbf{v}$	no$-\mathbf{z}$	no$-\mathbf{xy}$
Acc. [%]	82.08	79.66	65.89	66.53	65.52

Table 2.5: TCPCN accuracy (no Alg. 2.1 and Alg. 2.2) on 8 single targets using: all the point-cloud features in $\mathbf{p}_r$ (all), selectively leaving out the received power information (no-P), the velocity (no-v), the z coordinate (no-z) or the $x - y$ coordinates (no-xy)
.

Model	Training time [min]	No. of parameters
TCPCN (Ours)	13	$153,711$
PN + Gated Recurrent Unit (GRU)	19	$218,115$
mm-GaitNet [12]	32	$178,595$
bi-LSTM [8]	63	$3,237,379$

Table 2.6: Comparison between TCPCN and other models from the literature in terms of training time and number of parameters.

detector), by t_c the time to transmit the data using the UART port, by t_t the execution time of the DBSCAN clustering algorithm, the CM-KF tracking step and the data association, and by t_i the inference time of the classifier. We found that while t_p is stable and strictly lower than 10 ms, t_c is highly variable, mostly because of the variable number of detected points in the scene, and ranges between 0 ms (when no points are detected) and 25 ms (with 3 subjects). The clustering and tracking take on average $t_t = 12$ ms with 3 subjects, with very low variance. Being the radar frame duration $\Delta t \approx 67$ ms, the identification step has meet the inequality $t_i < \Delta t - \max t_p - \max t_c - t_t \approx 20$ ms. In the next section, we present a comparison between the proposed approach and two works from the literature in terms of accuracy and inference time, taking these considerations into account.

2.5.8 Comparison with state-of-the-art solutions

Out of the two other approaches from the literature (see Section 2.2), [12], does not obtain good results when subjects move freely, as neither a robust tracking method is implemented nor the identification information is used to improve the tracking performance, while [8] performs the identification in an *offline* fashion. In addition, they use different datasets. For these reasons, we chose to implement the classifiers from [12] and [8] and evaluate them on our multi-target test dataset using $K = 30$ input time steps and the same training data. As a baseline, we consider a model similar to TCPCN, but using a Recurrent Neural Network (RNN) instead of temporal convolutions after the point-cloud feature extraction block. We refer to this model as PN + GRU in the following, as it is obtained combining a feature extraction block similar to PointNet with a GRU layer [60], which is capable of learning long-term dependencies. GRU cells maintain a hidden state across time, processing it together with the current input vector to learn temporal features in the input sequence (see [60] for a detailed description of GRU cells). In our implementation,

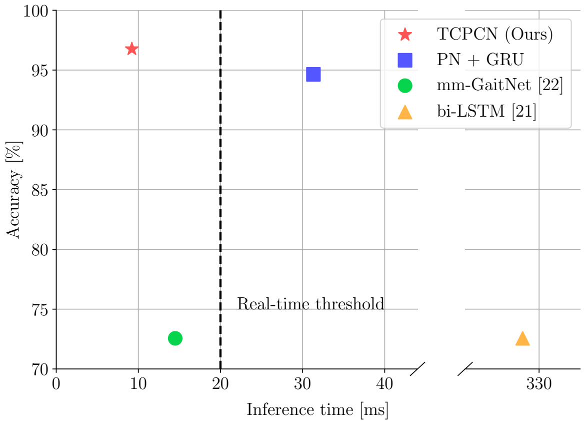


Figure 2.8: Performance comparison of the proposed TCPCN model against mm-GaitNet [12] and the bidirectional LSTM from [8]. As a baseline, we also evaluate a network similar to TCPCN that uses a GRU layer (PN + GRU) instead of temporal convolutions.

we use a GRU layer with 128 hidden units.

In Tab. 2.6, we compare the learning models in terms of training time and number of parameters. This evaluation has been conducted on an NVIDIA RTX 2080 GPU for all the models. The training time is affected by the processing speed of each NN model and by the convergence time of the training process (number of training epochs). We note that the processing time of convolutional models (TCPCN and mm-GaitNet) is lower than that of recurrent ones (PN + GRU and bi-LSTM). However, training is significantly faster for the two models featuring the proposed point-cloud feature extractor (TCPCN and PN + GRU) due to faster convergence.

A comparison of accuracy and inference time, measured on the NVIDIA Jetson board, is presented in Fig. 2.8. The most accurate models in identifying the subjects are our TCPCN and PN + GRU. This shows the superiority of using a point-cloud feature extractor, due to its invariance to the ordering of the input points. TCPCN proves to be slightly better than PN + GRU, meaning that dilated temporal convolutions do not only improve the inference and training times but are also more effective in extracting temporal features. Through a vertical dashed line, we mark the maximum inference time for the algorithms to run in real-time on the Jetson device, i.e., 20 ms (see Section 2.5.7): only two models satisfy this constraint, namely the proposed TCPCN and mm-GaitNet [12], which both exploit convolutions, as opposed to the RNN-based PN + GRU and bi-LSTM. In particular, TCPCN is the fastest model in making predictions, with an average inference time of 9.21 ± 2.12 ms.

2.6 Concluding remarks

In this chapter, we proposed a novel system that performs real-time person tracking and identification on an edge computing device using sparse point-cloud data obtained from a low-cost mmWave radar sensor. The raw signal undergoes several processing steps, including detection, clustering and Kalman filtering for position and subject extension estimation in the $x - y$ plane, followed by a fast neural network classifier based on a point-cloud specific feature extractor and dilated temporal convolutions. Our system significantly outperforms previous solutions from the literature, both in terms of accuracy and inference time, being able to reliably run in real-time at 15 fps on an NVIDIA Jetson TX2 board, identifying up to three subject among a group of eight with an accuracy of almost 92%, while simultaneously moving in an unseen indoor environment.

3

Contact Tracing and Temperature Screening via mmWave and Infrared Sensing

3.1 Introduction

This chapter moves beyond pure mmWave radar sensing, to show how person tracking and identification using RF signals can be paired with other sensing devices to provide diverse information. We tackle the problem of jointly performing unobtrusive elevated skin temperature screening and privacy preserving contact tracing in indoor environments.

Lately, *social distancing* has become a primary strategy to counteract the COVID-19 infection. Many research works [61], [62] have shown that it is an effective non-pharmacological approach and an important inhibitor for limiting the transmission of many contagious diseases such as H1N1, SARS, and COVID-19. Along with social distancing, *elevated skin temperature detection* and *contact tracing* have proven to be key to effectively contain the pandemic [63]. However, available methods to enforce these countermeasures often rely on RGB cameras and/or apps that need to be installed and continuously run on people's smartphones, often rising privacy concerns [64]. Moreover, currently adopted methods to screen people's temperature require individuals to stand in front of a thermal sensor, which may be impractical in heavily frequented public places. To this end, we propose milliTRACE-IR, a joint mmWave radar and infrared imaging sensing system that performs privacy preserving human body temperature screening and contact tracing in indoor spaces (see Fig. 3.1). Its main components are discussed next, emphasizing their novel aspects and the joint processing of the acquired sensor data.

On the one hand, the radar analyzes the reflections of a transmitted mmWave signal off the individuals that move in the monitored environment, returning *sparse point-clouds* that carry information about the subjects' locations and the velocity of their body parts. A novel point-

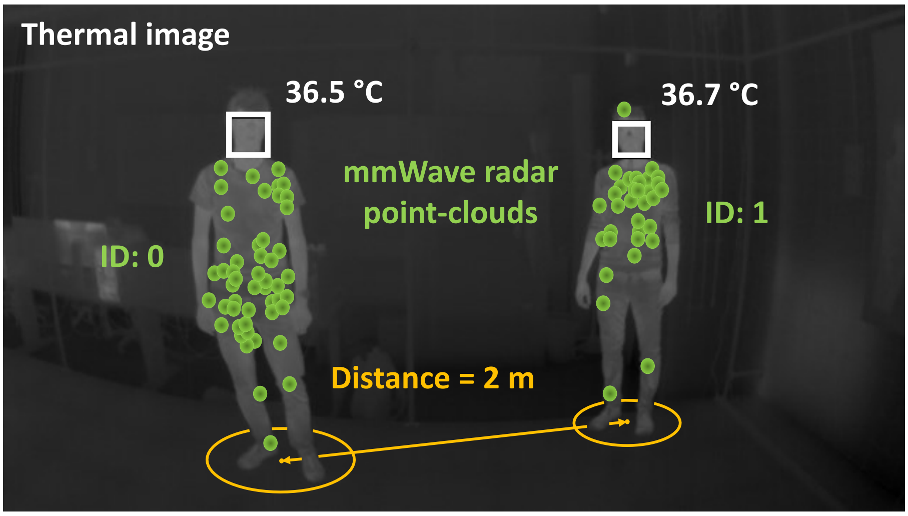


Figure 3.1: milliTRACE-IR performs body temperature screening and interpersonal distance estimation via sensor fusion of an infra-red thermal camera and mmWave radars. Individual gait features contained in the mmWave reflections enable contact tracing across different rooms.

cloud clustering method is designed, combining Gaussian Mixtures (GMs) [65] and DBSCAN [45], to distinguish the mmWave radio reflections from the subjects, as they move as close as 0.2 m to one another. The so obtained point-cloud clusters are used to track the subjects' positions in the physical space by means of a KF [42], and to obtain their gait-related features through a deep-learning based feature extractor. Finally, a novel person re-identification algorithm is proposed by exploiting Weighted Extreme Learning Machines (WELM).

On the other hand, the infrared imaging system, or Thermal Camera (TC), returns images whose pixels contain information on the *temperature of the objects* in the TC Field-of-View (FoV). To measure the subjects' temperature, at first, You Only Look Once v3 (YOLOv3) [66] is used to perform face detection in the TC images, by bounding those areas containing a human face. Hence, the obtained bounding boxes are tracked through an Extended Kalman Filter (EKF) [67] and the subjects' temperature is estimated by accumulating readings for each EKF track, according to a dedicated estimation and correction procedure. Through the EKF, the subject's distance from the TC is also estimated from the size of the corresponding bounding box by considering the non-linear part of the EKF, which is approximated by fitting a function over a set of experimental data points.

Tracks in the radar reference systems are associated with those in the TC image plane via an original algorithm that finds optimal matches for the readings taken by the two sensors, through their *joint* analysis. This makes it possible to take temperature measurements from a subject and reliably associate them with the highly precise tracking of his/her movement performed by the radar. In addition, the joint analysis of radar and TC data allows refining the temperature estimated through the TC: to mitigate the influence of the distance on the temperature readings [68], a regression function that provides temperature correction coefficients is fit from training data.

The final temperatures are obtained using such function with the accurate distances retrieved from the radar.

Hence, once a subject's temperature is measured, it is associated with the corresponding radar track and the subjects' movements and contacts inside the building are accurately monitored, by re-identifying the subjects as they move across the FoV of different radar devices. To the best of the author's knowledge, milliTRACE-IR is the first system that achieves temperature screening and human tracking through the joint analysis of radar and TC signals. Furthermore, it concurrently performs body temperature screening and contact tracing, while these aspects have been previously dealt with separately. A sensible usage model for the system is as follows: the TCs shall be deployed in strategic locations to allow an effective temperature screening, such as facing the building/room entrance, to ensure that people's faces are seen frontally for a reasonable amount of time, and that their TC images are only taken when they enter or leave the building/room. On the other hand, the radar can be utilized to track the subjects while moving inside the monitored indoor space. This ensures higher privacy with respect to RGB cameras.

The main contributions presented in this chapter are:

1. milliTRACE-IR, a joint mmWave radar and infrared imaging sensing system that performs unobtrusive and privacy preserving human body temperature screening and contact tracing in indoor spaces is designed and validated through an extensive experimental campaign.

2. A novel *data association* method is put forward to robustly associate tracks obtained from the mmWave radar and from the TC, where the radar returns the people coordinates in the physical space and the TC identifies people's faces in the thermal image space. The achieved precision and recall in the associations are as high as 97%.

3. An original *clustering algorithm for mmWave point-clouds* is devised, making it possible to resolve the radar reflections from subjects as close as 0.2 m.

4. A new WELM based *person re-identification* procedure is presented. The WELM is trained at runtime on previously unseen subjects, achieving an accuracy of 95% over six subjects with only 3 minutes of training data.

5. A novel method is designed to perform *elevated skin temperature screening* as people move freely within the FoV of the TC, without requiring them to stop and stand in front of the thermal sensor. For this, a dedicated approach is presented to mitigate the distortion in the TC temperature readings as a function of the distance, by also leveraging the accurate distance measures from the radar. Through this method, worst-case errors of 0.5 °C are obtained.

The chapter is organized as follows. In Section 3.2, the related work is discussed. Section 3.3 introduces some basic concepts about mmWave radars and thermal imaging systems, while in Section 3.4 the proposed approach is thoroughly presented. In Section 3.5.1, the implementation of milliTRACE-IR is described, while Section 3.5 contains an in depth evaluation of milliTRACE-IR on a real experimentation setup. Concluding remarks are provided in Section 3.6.

3.2 Related Work

In the literature, almost no work has focused on a joint approach to social distancing and people's body temperature monitoring which preserves the privacy of the users. Here, several prior works in related areas are discussed, highlighting the differences with respect to the proposed system.

Social distancing monitoring: Social distancing has been one of the most widely employed countermeasures to contagious diseases outbreaks [61]. Real-time monitoring of the distance between people in workplaces or public buildings is key for risk assessment and to prevent the formation of crowds. Existing approaches use either wireless technology like Bluetooth or Wi-Fi [69], [70], which require the users to carry a mobile device, or camera-based systems [71], which are privacy invasive. Other approaches use the Received Signal Strength Indication (RSSI) from cellular communication protocols [61] or wearables [72], although these are often inaccurate, especially when used in crowded places [61]. A lot of effort has been put into designing person detection and tracking algorithms for crowd monitoring and people counting [73] by using fixed surveillance cameras and mobile robots [74]. The main drawbacks of these methods are the intrinsic difficulty in estimating the distance between people from images or videos, along with the fact that the users have to be continuously filmed during their daily lives, which raises privacy concerns.

Concurrently, a large body of work has focused on Ultra Wide-Band (UWB) transmission for people tracking [8], [75], e.g., using mmWave radars, as these naturally allow measuring distances with decimeter-level accuracy. However, none of these works has tackled the problem of estimating interpersonal distances when people are very close to one another for extended periods of time; this is especially difficult with radio signals, as the separation of the reflections from different subjects becomes challenging.

Passive temperature screening: Infrared thermography is widely adopted for non-contact temperature screening of people in public places [76]. Due to the COVID-19 pandemic, there has been a growing interest in developing screening methods to measure the temperature of multiple subjects simultaneously, without requiring them to collaborate and/or to carry dedicated devices [77]. Approaches that involve the use of RGB cameras, e.g., [78], share the aforementioned privacy-related limitations.

The authors of [68] developed a Bayesian framework to measure the body temperature of multiple users using low-cost passive infrared sensors. The distance from the sensors and the number of subjects is also obtained. However, the working range of this system is very short (around 1.5 m for precise temperature estimation), so it is deemed inapt for monitoring a large indoor area.

Radar-thermal imaging association and fusion: Sensor fusion between radars and RGB cameras has been extensively investigated, see, e.g., [79], [80], while the joint processing of mmWave radar data and infrared thermal images was marginally treated [81]. In addition, the last paper only deals with the detection of humans using thermal imaging and does not address body temperature screening.

The present chapter is focused on the *data association* between a thermal camera and a mmWave radar over short periods of time, using the accurate radar distance estimates to refine the temperature reading. This makes it possible to consider scenarios where the thermal camera only covers a small portion of the environment (e.g., the entrance) so as to preserve the subjects' privacy, while a mmWave radar network can effectively monitor the whole indoor space.

mmWave radar person re-identification (Re-Id): RF based person Re-Id is a recent research topic. So far, many works have focused on person identification [7], [12], where the subjects to identify have been previously seen by the system, typically via a preliminary training phase. Re-Id is more challenging, as it addresses the recognition of *unseen* subjects, for which only a few radio samples are collected during system operation. Differently from camera image based Re-Id methods [16], RF approaches need to profile the users across time intervals of a few seconds, to extract robust person specific features [82]. To the best of the author's knowledge, only two works have proposed solutions to this problems [82], [83]. In both cases, a deep learning method trained on a large set of users is used to extract features from the human gait. At test time, the features obtained from the subjects to be re-identified are compared against those of a set of known individuals using distance-based similarity scores. This approach entirely depends on the feature extraction process, and the classifier does not learn to refine its decisions at *runtime*, as new samples become available. This is a weakness, as the gait features extracted from mmWave radars are known to be variable, e.g., across different days [6]. Conversely, milliTRACE-IR combines deep feature extraction with fast classifiers which are continuously trained and refined as new data is collected; this improves the robustness of the identification task.

3.3 Preliminaries

In this section we briefly recall the mmWave radar person detection method developed in the previous chapter, which we will reuse as the basic building block for our subsequent sensor fusion algorithms. Moreover, the main working principles of infrared thermal cameras are presented.

3.3.1 mmWave FMCW Radar

As thoroughly discussed in Chapter 2, a MIMO FMCW radar allows the joint estimation of the distance, the radial velocity and the angular position of the targets with respect to the radar device [11]. It works by transmitting sequences of *chirp* signals, linearly sweeping a bandwidth B, and analyzing their copies, which are reflected back from the environment. A full chirp sequence, termed *radar frame*, is repeated with period Δ seconds.

Distance, velocity and angle estimation

By computing the frequency shift induced by the delay of each reflection, the radar allows obtaining the distance and velocity of the targets with high accuracy. The use of multiple receiving antennas, organized in a *planar array*, allows obtaining the AoA of the reflections along the azimuth and the elevation dimensions, leveraging the different frequency shifts measured by the different antenna elements. This enables the localization of the targets in the physical space.

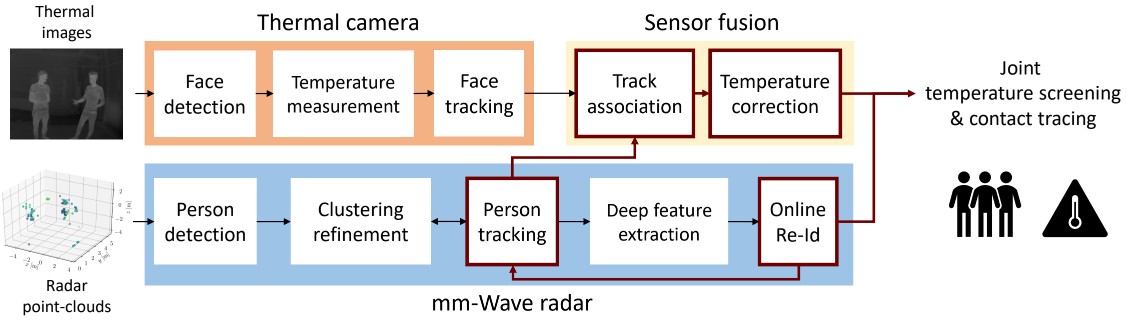


Figure 3.2: milliTRACE-IR signal processing workflow.

Radar detection

The raw output of the radar is typically high dimensional for mmWave devices, due to the high resolution. To sparsify the signal and perform a detection of the main reflecting points, a typical approach is the Constant False Alarm Rate (CFAR) algorithm [43], which consists of applying a dynamic threshold on the power spectrum of the output signal. A further processing step is required to remove the reflections from static objects, i.e., the *clutter*. This operation is performed using a MTI high pass filter that removes the reflections with Doppler frequency values close to zero [43].

Radar point-clouds

After the detection phase, a human presence in the environment typically generates a large number of detected points. This set of points, usually termed radar *point-cloud*, can be transformed into the 3-dimensional Cartesian space $(x - y - z)$ using the distance, azimuth and elevation angles information of the multiple body parts. In addition, the velocity of each point is also retrieved, along with the strength of the corresponding signal reflection.

In the following, the point-cloud outputted by the radar at frame k is referred to as $\mathcal{P}_k$, containing a variable number of reflecting points. Each point, $\mathbf{p} \in \mathcal{P}_k$, is described by vector $\mathbf{p} = \left[x, y, z, v, P^{\text{RX}}\right]^T$, including its coordinates x, y, z, its velocity v and reflected power P^{RX}.

3.3.2 Infrared Thermal Cameras

Infrared thermal imaging deals with detecting radiation in the long-infrared range of the electromagnetic spectrum ($\sim 8 - 15$ μm) and producing images of that radiation, called *thermograms*. According to the *Planck's Law*, infrared radiation is emitted by all objects with temperature $T > 0$ K [84]. Since the radiation energy emitted by an object is positively correlated to its temperature, from the analysis of the received radiation it is possible to measure the object's temperature.

A thermographic camera, or *thermal camera*, is a device that is capable of creating images of the detected infrared radiation. The operating principle is quite similar to that of a standard camera, and the same relations described by the so-called *pinhole camera model* hold [85]. Within this approximation, the coordinates of a point $\mathbf{a} = [a_x, a_y, a_z]^T$ in the three-dimensional space are projected onto the image plane of an ideal pinhole camera through a very small aperture. Mathematically, this operation is described as $\mathbf{a}^{\text{proj}} = \mathbf{\Psi a}$, where $\mathbf{a}^{\text{proj}}$ is the projected point and $\mathbf{\Psi}$ is the intrinsic matrix of the camera that contains information about its focal lengths, pixel dimensions and position of the image plane. However, when dealing with a real thermal camera, this approximation may be insufficient and the *radial* and/or *tangential* distortions introduced by the use of a lens and by inaccuracies in the manufacturing process may additionally have to be accounted for. On the image plane, an array of infrared detectors is responsible for measuring the received radiation, which is sampled and quantized to produce digital information. The pixels of the final image that is returned by a thermal camera contain information about the temperature of the corresponding body/object part, encoded into the pixel intensity.

3.4 Proposed Approach

This chapter considers the problem of monitoring an indoor environment covered by multiple mmWave radar sensors, which span over different rooms and corridors. A few infrared thermal cameras are placed at strategic locations to perform accurate temperature screening of the people in the indoor space without compromising their privacy, e.g., at the building's entrance.

From a high-level perspective, milliTRACE-IR performs the following operations.

(1) **Person detection and temperature measurement:** When people enter the monitored indoor space, the system concurrently performs face detection from the infrared images captured by the thermal camera and person detection using the mmWave radar point clouds.

1. From the Thermal Camera (TC) images, a face detector is used to obtain bounding boxes enclosing the faces of the detected subjects, (Section 3.4.2). A measure of their body temperature is obtained from the intensity of the thermal image pixels in the bounding box, see Section 3.4.3. While milliTRACE-IR works independently of the specific face detector architecture used, in the implementation YOLOv3 is used [66].

2. Concurrently, radar signal processing is used to detect and group the point-clouds from different subjects and estimate their positions (Section 3.4.4). A novel clustering algorithm based on DBSCAN and GM models is put forward to separate the contributions of closeby subjects (Section 3.4.5).

(2) **Radar-TC person tracking:** a KF is independently applied to the TC images and to the radar point-clouds, following the approach presented in Chapter 2, to respectively track the subjects' movements within the thermal images and in the indoor Cartesian space. Standard KF based tracking in the thermal image plane is here modified to achieve a coarse estimation of the distance of the subjects, based on the dimension of their face bounding box Section 3.4.2. In this phase, each subject track is associated with a unique numerical identifier.

(3) **Radar-TC track association:** As a subject exits the FoV of the TC, his/her body temperature is associated with the corresponding trajectory from the mmWave radar, by performing

a track-to-track association between TC tracks and radar tracks. This association algorithm is based on the subjects' distances from the TC, and on the radar estimated positions of the subjects, projected onto the thermal image plane (Section 3.4.6). After the association, the temperature measurement is corrected accounting for the distance of each person from the TC, using the more precise distance estimates provided by the radar, Section 3.4.3.

(4) **Radar-based person re-identification:** During the radar tracking process, the point-cloud sequences generated by each subject are collected and fed to a deep neural network that performs gait feature extraction (Section 3.4.8). The resulting gait features are organized into a labeled training set, where labels are obtained from the track identifiers. When a subject exits the FoV of a radar and enters that of another radar placed in a different room or corridor, a WELM based classifier [86] is trained on-the-fly and used to re-identify the subject at runtime (Section 3.4.10). This robust and lightweight person Re-Id process, based on the gait features extracted from the radar point-clouds, enables contact tracing across large indoor environments.

3.4.1 Notation

The system operates at discrete time-steps, $k = 1, 2, \dots,$ each with fixed duration of Δ seconds, also referred to as *frame* in the following. Boldface, capital letters refer to matrices, e.g., $\mathbf{X}$, with elements X_{ij}, whereas boldface lowercase letters refer to vectors, e.g., $\mathbf{x}$. $\mathbf{X}^{-1}$ denotes the inverse of matrix $\mathbf{X}$, and $\mathbf{x}^T$ denotes the transpose of vector $\mathbf{x}$. $\mathbf{x}_k$ refers to vector $\mathbf{x}$ at time k, x_j refers to element j of $\mathbf{x}$ and $(\mathbf{x}_k)_j$ is element j of $\mathbf{x}_k$. $\mathcal{N}(\mu, \sigma^2)$ indicates a Gaussian random variable with mean μ and variance σ^2. Notation $||\mathbf{x}||_2$ indicates the Euclidean norm of vector $\mathbf{x}$, while $||\mathbf{x}||_{\mathbf{\Gamma}}= \sqrt{\mathbf{x}^T\mathbf{\Gamma}\mathbf{x}}$ denotes the norm induced by matrix $\mathbf{\Gamma}$. The diagonal matrix with elements $x_1, x_2, \dots, x_n$ is denoted by diag $[x_1, x_2, \dots, x_n]$. $|\mathcal{X}|$ indicates the cardinality of set $\mathcal{X}$ while $\log(\cdot)$ denotes the natural logarithm.

3.4.2 Thermal Camera: Face Detection and Tracking

The detection of the subjects in the thermal camera images is performed by means of a face detector that computes rectangular bounding boxes delimiting the faces of the people within the FoV. The bounding boxes are used to track the positions of the subjects in the subsequent instants and to identify a Region Of Interest (ROI) from which the temperature of the targets is obtained. milliTRACE-IR is independent of the particular face detector used, provided that it outputs bounding boxes enclosing the faces of the subjects. In the implementation, YOLOv3 [66] is used due to its excellent performance in terms of accuracy and speed.

To track the faces of the subjects in the image plane, an EKF is employed [67]. Define the *state* vector of a target subject at time k, as $\mathbf{x}_k = [x^c_k, y^c_k, \dot{x}^c_k, \dot{y}^c_k, h_k, d_k, \dot{d}_k]^T$, where x^c_k, y^c_k are the true coordinates of the center of his/her face in the thermal image, $\dot{x}^c_k, \dot{y}^c_k$ its velocities along the vertical and horizontal directions, h_k is the true height of the bounding box enclosing the subject's face, d_k, the distance of the target from the camera in the physical space, and $\dot{d}_k$ its time derivative (rate of variation).

The observation vector obtained from the YOLOv3 face detector, denoted by $\mathbf{z}_k = [\tilde{x}_k^c, \tilde{y}_k^c, \tilde{h}_k]^T$, contains noisy measurements of the face position and height (represented by the height of the bounding box), which are distinguished from their true values by the superscript "$\tilde{\ }$". Denote the observation noise by vector $\mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, with $\mathbf{R} = \text{diag}\left(\sigma^2_{\tilde{x}^c}, \sigma^2_{\tilde{y}^c}, \sigma^2_{\tilde{h}}\right)$, with diagonal elements representing the (constant) observation noise variances of $\tilde{x}_k^c$, $\tilde{y}_k^c$ and $\tilde{h}_k$, respectively. In the implementation $\sigma^2_{\tilde{x}^c} = \sigma^2_{\tilde{y}^c} = 0.01$ and $\sigma^2_{\tilde{h}} = 20$ are used.

The EKF state transition model is defined as $\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k)$, where $f(\cdot)$ is the transition function, connecting the system state at time k, $\mathbf{x}_k$, to that at time $k+1$, $\mathbf{x}_{k+1}$, and vector $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ represents the process noise. In the model used in this chapter, the process noise includes 4 independent components, representing two random accelerations of the bounding-box center coordinates, u_k^x, u_k^y, a random noise term for the bounding-box dimension, u_k^h, and a random acceleration for the subject's distance, u_k^d. Therefore, it can be written $\mathbf{u}_k = \left[u_k^x, u_k^y, u_k^h, u_k^d\right]^T$ with covariance matrix $\mathbf{Q} = \text{diag}\left[\sigma_x^2, \sigma_y^2, \sigma_h^2, \sigma_d^2\right]$. In the implementation, $\sigma_x^2 = \sigma_y^2 = \sigma_d^2 = 5$ and $\sigma_h = 5.148$ are used (see Section 3.4.2).

Assuming that the target moves according to a *constant velocity* (CV) model, from the state definition it follows that

$$f(\mathbf{x}_k, \mathbf{u}_k) = \begin{bmatrix} x_k + \Delta\dot{x}_k + u_k^x\Delta^2/2 \\ y_k + \Delta\dot{y}_k + u_k^y\Delta^2/2 \\ \dot{x}_k + u_k^x\Delta \\ \dot{y}_k + u_k^y\Delta \\ g\left(d_k + \Delta\dot{d}_k + u_k^d\Delta^2/2\right) + u_k^h \\ d_k + \Delta\dot{d}_k + u_k^d\Delta^2/2 \\ \dot{d}_k + u_k^d\Delta \end{bmatrix}, \tag{3.1}$$

where the only non-linear term is function $g(\cdot)$, which relates the subject's distance extracted by the thermal camera to the height h_k of the bounding-box enclosing his/her face. The proposed approach consists in *(i)* obtaining an estimate for $g(\cdot)$ in an *offline* fashion using training data, and *(ii)* using such estimate in the EKF model. These two steps are detailed next.

Estimation of function $g(\cdot)$

Function $g(\cdot)$ maps the distance of the target from the thermal camera d_k, at time k, onto the corresponding height of the bounding box, h_k, as follows,

$$h_k = g(d_k) + u_k^h. \tag{3.2}$$

Using N_t training samples $\{h_i, d_i\}_{i=1}^{N_t}$ containing the true distances of the target, d_i, and the measured bounding box height, h_i, $g(\cdot)$ is obtained solving an *offline* non-linear least-squares (LS) problem of the form

$$\underset{g}{\text{argmin}} \sum_{i=1}^{N_t} \left(h_i - g(d_i)\right)^2. \tag{3.3}$$

From the equations of the pinhole camera model [85], $g(\cdot)$ is restricted to the family of hyperbolic functions with shape $g(d_i) = b_0/(d_i + b_1) + b_2$, reducing the problem to that of estimating the parameters b_0, b_1, and b_2, i.e.,

$$\underset{b_0, b_1, b_2}{\operatorname{argmin}} \sum_{i=1}^{N_t} \left(h_i - \frac{b_0}{d_i + b_1} + b_2 \right)^2 . \tag{3.4}$$

This optimization problem is here solved using the Levenberg-Marquardt algorithm [87] for non-linear LS fitting: with the experimental setup used in this chapter, $b_0 = 162.04, b_1 = 0.61, b_2 = -14.79$ are obtained.

Note that the process noise acts on the bounding-box dimension in two ways, inside the function $g(\cdot)$, modeling the uncertainty in the subject's distance due to the random acceleration, and through the additive term u_k^h, modeling the imperfect estimation of $g(\cdot)$ itself. The variance of u_k^h can be estimated from the residuals, after fitting the training measurements with function $g(\cdot)$.

Using $g(\cdot)$ in the EKF

Due to the non-linear dependence of the state $\mathbf{x}_k$ on the process noise $\mathbf{u}_k$, in the EKF operations the following transformed process noise covariance matrix is used [88]

$$\mathbf{Q}'_k = \mathbf{L}_k \mathbf{Q} \mathbf{L}_k^T, \quad \text{with } \mathbf{L}_k = \left. \frac{\partial f\left(\mathbf{x}_k, \mathbf{u}_k\right)}{\partial \mathbf{u}_k} \right|_{\hat{\mathbf{x}}_{k|k}}, \tag{3.5}$$

where matrix $\mathbf{L}_k$ is the Jacobian of function $f(\cdot)$ with respect to the process noise vector, evaluated for the current state estimate. Using the above system model, the system state estimate at time k, $\hat{\mathbf{x}}_k$, is recursively obtained along with the corresponding error covariance matrix, $\mathbf{P}_k$. By definition of the EKF state, this allows us to get a coarse estimate of the distance of the subjects from the TC, which is exploited in the radar-TC data association step, see Section 3.4.6.

3.4.3 Thermal Camera: Subject Temperature Estimation

The body temperature is obtained from the thermal camera readings in the bounding-boxes contained in $\hat{\mathbf{x}}_k$, for each subject, and for all the time steps in which they are tracked by the EKF. At any given time k, a single (noisy) temperature measurement, $\tilde{T}_k$, is extracted by taking the maximum value across all the pixels in the current bounding box. Denoting by B_k the 2-D region of the image enclosed by the bounding box, and by B_{ki} the intensity of its pixel i, it holds $\tilde{T}_k = \max_i B_{ki}$.

3.4.4 mmWave Radar: People Detection and Tracking

As presented in Chapter 2, our approach to multiperson tracking from sparse mmWave radar point clouds includes

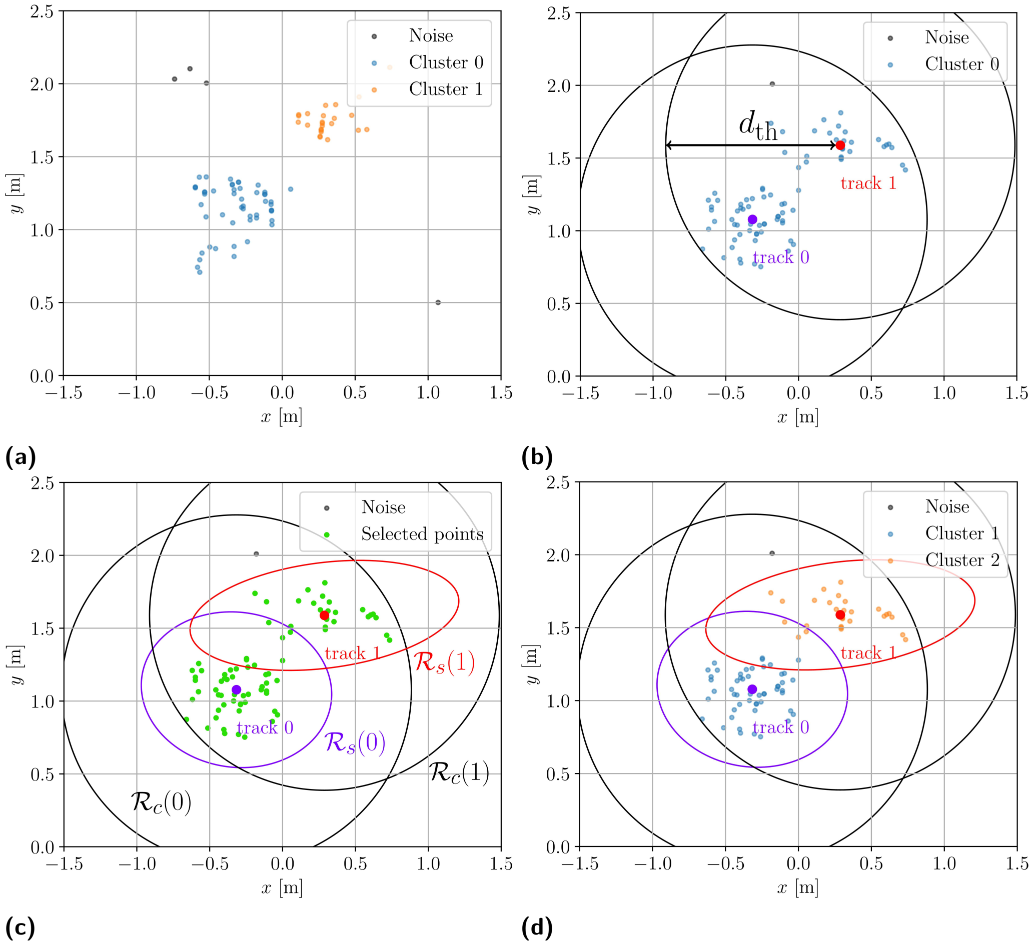


Figure 3.3: Illustration of the proposed clustering method. In (a) the point-clouds belonging to 2 subjects are well separated and DBSCAN outputs the correct clustering. In the next time-step, (b), DBSCAN fails and merges the two clusters into one. The proposed method selects the points to re-cluster using the tracks positions together with Eq. (3.6) and Eq. (3.7), as shown in (c), and outputs the correct result using GM on the selected points with $n_{\mathcal{G}} = 2$, see (d).

(i) **detection**: using density-based clustering to separate the points generated by the subjects from clutter and noise;

(ii) **tracking**: applying Kalman filtering techniques [42] on each cluster centroid to track the movement trajectory of each subject in space.

In the following, we assume that the same processing pipeline as in Chapter 2 is used, specifically applying DBSCAN [45], and KF based tracking. In the KF tracker, the *state* of each subject at time k is defined as $\mathbf{s}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$, containing the $x - y$ subject's coordinates and the corresponding velocities. The state evolution is assumed to obey $\mathbf{s}_k = \mathbf{A}\mathbf{s}_{k-1}$, where the transition matrix $\mathbf{A}$ represents a Constant Velocity (CV) model [46]. The KF computes an estimate of the state for a target subject at time k, denoted by $\hat{\mathbf{s}}_k$, by sequentially updating the predictions from the CV model with the new observations. The association between the new observations (time

k) and the previous states (time $k-1$) exploits the Nearest-Neighbors Cheap Joint Probabilistic Data Association (NN-CJPDA) [46], [89].

We stress that DBSCAN has proven to be robust and accurate as long as the subjects do not come too close to one another [8], [12], [46], see also Fig. 3.3a. When this occurs (Fig. 3.3b), the algorithm often fails to distinguish between adjacent subjects, merging their contributions into a single cluster [90]. For this reason, milliTRACE-IR improves the previous DBSCAN and KF based signal processing pipeline with a novel clustering procedure to better resolve the point-clouds of subjects that are close to one another. The designed solution to enhance the tracking accuracy in such cases is a major contribution of the present chapter and is detailed next.

3.4.5 mmWave Radar: Highly Accurate Clustering

As a possible solution to DBSCAN drawbacks, one may adjust the parameters ε and m_{pts} so as to correctly resolve the clustering ambiguity, even for closely spaced targets. However, ε and m_{pts} interact in a complex and often unpredictable way, making the design of such adaptation rule difficult.

milliTRACE-IR adopts a different approach, which combines *(i)* the standard DBSCAN algorithm with fixed ε and m_{pts}, *(ii)* the spatial locations of the subjects, available from the tracking procedure, and *(iii)* the Gaussian Mixture (GM) clustering algorithm [65]. The designed algorithm, reported in Alg. 3.1 and exemplified in Fig. 3.3, proceeds as follows. At first, the DBSCAN algorithm is applied to obtain an estimate of the clusters and a reasonable separation between the noise points and those belonging to actual subjects, using $\varepsilon = 0.4$ m and $m_{\text{pts}} = 10$. DBSCAN outputs a cluster label for each point $\mathbf{p} \in \mathcal{P}_k$, denoted by $\ell_{\mathbf{p}}$. Clusters are denoted by $\mathcal{C}_n$, and their centroids by $\bar{\mathbf{c}}_n$, with $n = 1, \dots, n_k$.

The next step is to identify which of the tracked subjects get closer than a critical distance d_{th} from one another. The clusters provided by standard DBSCAN for these subjects are expected to be incorrect, as the point-cloud data from these would be merged into a single cluster. To pinpoint these subjects, their KF state is leveraged, which corresponds to a filtered representation of their trajectories. Consider track t at time k, its coordinates are predicted as $\hat{\mathbf{s}}_k^t = \mathbf{A}\hat{\mathbf{s}}_{k-1}^t$ (see line $2-3$ in Alg. 3.1). For any two subjects with associated tracks t and t', milliTRACE-IR checks whether $||\hat{\mathbf{s}}_k^t - \hat{\mathbf{s}}_k^{t'}||_2 < d_{\text{th}}$. If this occurs, as shown in the example of Fig. 3.3b for tracks $t = 0$ and $t' = 1$, t and t' are termed *nearby subjects*. Hence, define $\mathcal{G}$ as the set of subjects that are mutually within a radius of d_{th} from one another. A group $\mathcal{G}$ can be constructed starting from any subject and recursively adding all the subjects who are closer than d_{th} from any of the set members. If a subject has no other subjects within distance d_{th}, it will be the only member of his group. Collecting all the disjoint groups, constructed from the maintained tracks at time k, set $\boldsymbol{\mathcal{G}}_k(d_{\text{th}})$ is obtained, containing all the nearby subjects groups. Once the nearby groups are identified, the ambiguities inside each group $\mathcal{G}$ containing more than one member are resolved by recomputing the clustering labels as follows. Consider a single group $\mathcal{G}$. To delimit the region where the clustering has to be refined, the following additional regions are defined. The sample covariance matrix of the *last* cluster associated with track t is denoted by $\boldsymbol{\Sigma}_n^t$, and contains information about the shape of the

Algorithm 3.1 Clustering refinement method.

```
Input: States of the targets at time $k-1$, observed point-cloud at time $k$, $\mathcal{P}_k$.
Output: Labels $\ell_{\mathbf{p}}$, $\forall\, \mathbf{p} \in \mathcal{P}_k$.
 1: $\{\ell_{\mathbf{p}}\}_{\mathbf{p}\in\mathcal{P}_k}, \{\mathcal{C}_n\}_{n=1}^{n_k} \leftarrow \text{DBSCAN}(\varepsilon, m_{\text{pts}}, \mathcal{P}_k)$
 2: $\hat{\mathbf{s}}_k^t \leftarrow \mathbf{A}\hat{\mathbf{s}}_{k-1}^t$ all maintained tracks $t$
 3: Find groups of nearby subjects $\boldsymbol{\mathcal{G}}_k(d_{\text{th}})$
 4: for each $\mathcal{G} \in \boldsymbol{\mathcal{G}}_k(d_{\text{th}})$
 5:     $n_{\mathcal{G}} \leftarrow |\mathcal{G}|$
 6:     if $n_{\mathcal{G}} > 1$
 7:         $\mathcal{R}(\mathcal{G}) \leftarrow \bigcup_{t\in\mathcal{G}}(\mathcal{R}_c(t) \cap \mathcal{R}_s(t))$
 8:         $\mathcal{S} \leftarrow \{\mathbf{p} \in \mathcal{C}_n \text{ such that } \bar{\mathbf{c}}_n \in \mathcal{R}(\mathcal{G})\}$
 9:         discard $\ell_{\mathbf{p}}, \forall\, \mathbf{p} \in \mathcal{S}$
10:         $\{\ell_{\mathbf{p}}\}_{\mathbf{p}\in\mathcal{S}}, \{\pi_q\}_{q=1}^{n_{\mathcal{G}}} \leftarrow \text{GM}(n_{\mathcal{G}}, \mathcal{S})$
11:         discard cluster $q$ if $\pi_q < \pi_{\text{thr}}$
12:     end if
13: end for
```

subject's cluster. The regions of the plane containing the points that are within a radius of d_{th} from $\hat{\mathbf{s}}_k^t$, can be written as

$$\mathcal{R}_c(t) = \left\{\mathbf{x} \in \mathbb{R}^2 \text{ s.t. } \left|\left|\mathbf{x} - \hat{\mathbf{s}}_k^t\right|\right|_2 < d_{\text{th}}\right\}, \tag{3.6}$$

and the regions of points with a squared Mahalanobis distance smaller than γ are

$$\mathcal{R}_s(t) = \left\{\mathbf{x} \in \mathbb{R}^2 \text{ s.t. } \left|\left|\mathbf{x} - \hat{\mathbf{s}}_k^t\right|\right|^2_{(\mathbf{\Sigma}_n^t)^{-1}} < \gamma\right\}. \tag{3.7}$$

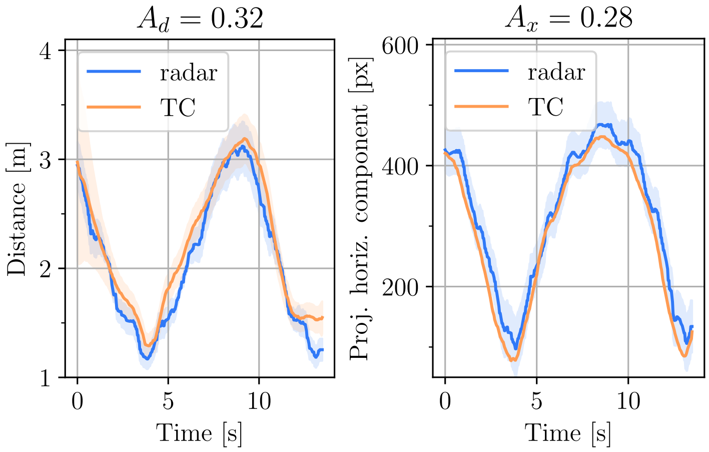
In the implementation, $d_{\text{th}} = 1.2$ m and $\gamma = 9.21$ were used.* Then, the labels assigned by DBSCAN to all the points belonging to a cluster whose centroid falls inside region $\mathcal{R}(\mathcal{G}) = \cup_{t\in\mathcal{G}}(\mathcal{R}_c(t) \cap \mathcal{R}_s(t))$, are discarded (lines $7-9$ in Alg. 3.1).† This set of points is denoted by $\mathcal{S}$.

Then, the GM algorithm is applied to the points belonging to set $\mathcal{S}$ to refine the clusters within this region, see the green points in Fig. 3.3c. As GM requires the number of clusters to be specified in advance, it is set to be equal to the number of subjects in the group, i.e., $n_{\mathcal{G}} = |\mathcal{G}|$. The GM algorithm outputs the labels $\ell_{\mathbf{p}}$ for each point $\mathbf{p} \in \mathcal{S}$ and the weight of the Gaussian component associated with each GM cluster, $\pi_q \in [0,1], q = 1, \dots, n_{\mathcal{G}}$, with $\sum_q \pi_q = 1$. The new labels are used to replace the ones previously found by DBSCAN (Fig. 3.3d), unless the GM clusters have very small weights, i.e., the new clusters having $\pi_q < \pi_{\text{thr}}$ are discarded and treated as noise points. The threshold value used in the implementation is $\pi_{\text{thr}} = 0.1/n_{\mathcal{G}}$.

The proposed method effectively solves the problem faced by DBSCAN in resolving subjects close to one another. The cost of this improvement is that an additional GM algorithm has to be applied to a subset of the point-cloud, however, at each time k the number of points in this subset is typically much smaller than that in the full point-cloud $\mathcal{P}_k$.

*The value of γ corresponds to a probability of 99% of falling inside the region, assuming that the points in the cluster are distributed on the plane according to a Gaussian distribution around $\hat{\mathbf{s}}_k^t$.

†Discarding a label corresponds to setting it equal to that used by DBSCAN to represent noise points.



(a) **(b)**

Figure 3.4: Example of distance (a) and horizontal projection (b) estimates from a track. The shaded areas represent the standard deviations. The corresponding values for A_d and A_x are shown above.

3.4.6 Radar and Thermal Camera Data Association

Upon tracking the subjects in the TC image plane and in the physical space, respectively using the measurements from the TC and from the mmWave radar sensor, a track-to-track association method is applied to link the movement trajectory of each person to his/her body temperature.

Assume that, at time k, the system has access to N_k^{rad} tracks from the radar sensor and N_k^{tc} tracks from the thermal camera, indicized by i and j, respectively. The data association strategy used in milliTRACE-IR consists in *(i)* computing a *cost* for each association $(i \leftrightarrow j)$, and *(ii)* solving the resulting combinatorial cost minimization problem to associate the best matching track pairs. The main challenge in the association of radar and thermal camera tracks is the design of a cost function that grants robustness in the presence of multiple targets, which may enter the monitored area in unpredictable ways. The key point is to gauge the similarity of the tracks by comparing them in terms of common quantities, which can be estimated from both devices.

Assume also that the two sensors are located in the same position and with the same orientation (co-located). In this setup, *(i)* the *distance* between the subjects and the sensors is the same, so its estimate should match for tracks representing the same subjects, and *(ii)* the radar KF states containing the coordinates of the subjects' positions can be projected onto the TC image plane; after this operation, the horizontal component of the radar projections and the horizontal component of the TC bounding boxes position should match for correctly associated tracks. To reliably associate radar and TC tracks, milliTRACE-IR uses a cost function consisting of the following components, see also Fig. 3.5

Estimated distance cost. Denote by d_k^i the estimated distance of radar track i, and by d_k^j the estimated distance of TC track j. Recalling that $\hat{\mathbf{s}}_k^i$ is the position of subject i at time k, d_k^i is computed using Pythagora's formula as $d_k^i = \sqrt{\left(\hat{\mathbf{s}}_k^i\right)_1^2 + \left(\hat{\mathbf{s}}_k^i\right)_2^2}$. Distance d_k^j, instead, is retrieved directly from the tracking state of the TC. Considering K subsequent time steps where radar track i and TC track j are both available, the estimated distance cost is defined as

$$A_d(i,j) = \frac{1}{K} \sum_{k=1}^{K} \frac{(d_k^i - d_k^j)^2}{\sigma^2_{d_k^i} + \sigma^2_{d_k^j}}, \tag{3.8}$$

where $\sigma^2_{d_k^i}$ and $\sigma^2_{d_k^j}$ represent the variances of the two distance estimates. An illustrative example is shown in Fig. 3.4a.

Projected horizontal component cost. The horizontal components of the radar state projection and of the TC bounding box center are respectively denoted by x_k^i and x_k^j. The radar positions provided by the KF state have only two dimensions, x and y (the first and second components of the state vector). However, three-dimensional vectors are needed for their proper projection onto the TC image plane. For this reason, a 0-valued z component is artificially added, under the assumption that the subjects' position at height 0 is the one being tracked. For this, an augmented subject's position vector, $\mathbf{a}_k^i = [(\hat{\mathbf{s}}_k^i)_1, \left(\hat{\mathbf{s}}_k^i\right)_2, 0]^T$, is defined. x_k^i is computed by projecting the radar coordinates $\mathbf{a}_k^i$ onto the TC image plane, as $\mathbf{a}_k^{i,\text{proj}} = \mathbf{\Psi} \mathbf{a}_k^i$ (see Section 3.3.2), applying to it a radial distortion based on the estimated distortion coefficients and retaining only the x-axis component. Projection x_k^j corresponds to the x coordinate of the TC tracked state. The projected horizontal component cost is defined, for K subsequent time steps of radar track i and TC track j, as

$$A_x(i,j) = \frac{1}{K} \sum_{k=1}^{K} \frac{(x_k^i - x_k^j)^2}{\sigma^2_{x_k^i} + \sigma^2_{x_k^j}}, \tag{3.9}$$

where $\sigma^2_{x_k^i}$ and $\sigma^2_{x_k^j}$ are the variances of the two estimates. An illustrative example is shown in Fig. 3.4b.

Track length coefficient. Recalling that Δ is the (constant) sampling interval, the proposed cost function accounts for the length K of the tracks that are to be associated, favoring longer tracks. To this aim, the following coefficient is defined,

$$\rho(K) = \frac{1}{\log(K\Delta)}. \tag{3.10}$$

Note that $\rho(K)$ is a weight factor for a cost (see the later Eq. (3.11)), which decreases with the track length K. This means that a smaller cost is implied when the associated tracks i and j are longer. Also, in the implementation, it holds $K > 1/\Delta$, so $\rho(K)$ is always positive.

Association cost function for radar and TC tracks. The association cost $A(i,j)$ for the tracks pair (i,j) (i refers to a radar track and j to a TC track) is obtained summing Eq. (3.8) and Eq. (3.9), to gauge how well the two tracks match in terms of their estimated distance across time, and estimated position on the horizontal projected axis on the TC image plane, respectively. The

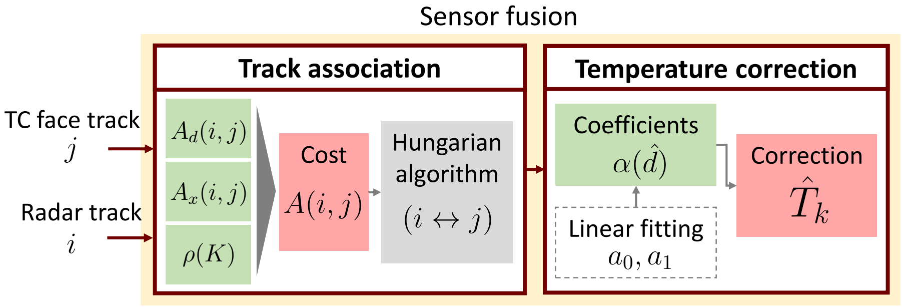


Figure 3.5: Block diagram of the sensor fusion step.

sum is then weighted by the coefficient of Eq. (3.10). Formally, $A(i,j)$ is given by

$$A(i,j) = \rho(K)\left[A_d(i,j) + A_x(i,j)\right]. \tag{3.11}$$

Costs $A(i,j)$, $i = 1, \dots, N_k^{\text{rad}}$, $j = 1, \dots, N_k^{\text{tc}}$, are arranged into an $N_k^{\text{rad}} \times N_k^{\text{tc}}$ matrix, and the optimal association of tracks is obtained by minimizing the overall cost, computed through the Hungarian algorithm [51]. The Hungarian algorithm takes the cost matrix as input and solves the problem of pairing each radar track with a single TC track (by minimizing the total cost), with an overall complexity of $O((N_k^{\text{rad}} N_k^{\text{tc}})^3)$.

In general, the radar and the TC would be deployed at different spatial locations. However, knowing their relative position and orientation, a roto-translation matrix $\mathbf{\Phi}$ can be obtained to geometrically transform the data into a new coordinate system where the TC and the radar sensors are co-located, as described above. In this chapter, the TC position and orientation are selected as the reference coordinate system, and the positions estimated from the radar sensor are transformed into it.

3.4.7 Temperature Correction

In line with [68], the direct reading of each subject's temperature , $\tilde{T}_k$, is subject to a scaling factor, $\alpha(d_k)$, with respect to the true temperature T, where $\alpha(d_k)$ depends on the distance from the TC, i.e.,

$$T = \alpha(d_k)\tilde{T}_k. \tag{3.12}$$

For an accurate temperature screening, the scaling factor $\alpha(d_k)$ is estimated from the training data, considering a linear model of the form

$$\alpha(d_k) = a_0 + a_1 d_k. \tag{3.13}$$

Table 3.1: Summary of the architecture and training parameters of the NN used for gait feature extraction.

Architecture	
Layer/block	**Size**
PC features [6]	$3+2$ shared MLPs, (98, 196)
Temporal conv. [6]	3 Conv. (3×3), $32, 64, 128$ filt.
Temporal conv. [6]	3 Conv. (3×3), $256, 128, 32$ filt.
Global average pooling	32
Fully connected	32
L_2 normalization	32
Fully connected	16
Training parameters	
Learning rate	10^{-4}
Optimizer	Adam [53]
Number of epochs	250
$\mathcal{L}_{\text{cen}}$ weight, ω	0.5
L_2-regularization parameter	8×10^{-5}
Dropout rate	0.4
Triplet margin, μ	1

Using $N_t^{'}$ training measurements $\{\tilde{T}_i, d_i, T\}_{i=1}^{N_t^{'}}$, the fitting coefficients a_0, a_1 are obtained by solving

$$\underset{a_0, a_1}{\operatorname{argmin}} \sum_{i=1}^{N_t^{'}} \left(T - \alpha(d_i)\tilde{T}_i\right)^2 . \tag{3.14}$$

From the above optimization problem, in this chapter the above parameters are set to $a_0 = 1.116$, $a_1 = 0.013$. At system operation time, denoting by M the number of time-steps for which the subject is correctly tracked by the EKF, his/her true temperature at time k is finally estimated as

$$\hat{T}_k = \frac{1}{M} \sum_{j=k-M+1}^{k} \alpha(\hat{d}_j)\tilde{T}_j, \tag{3.15}$$

where $\alpha(\cdot)$ is defined in Eq. (3.13), using the parameters obtained from Eq. (3.14), while $\hat{d}_j$ is an estimate of the distance obtained by the system at time-step j. To improve the temperature estimates, milliTRACE-IR performs sensor fusion by exploiting the association between the TC face tracks and the mmWave radar tracks (see Section 3.4.6). In Eq. (3.15), the coefficients $\alpha(\hat{d}_j)$ are computed using the distances estimated by the mmWave radar device, as these are much more accurate than those obtained from the TC. The impact of combining the temperature information from the TC and the accurate distance estimation capabilities of the radar is investigated in Section 3.5.3. The block diagram for the temperature correction step is shown in Fig. 3.5.

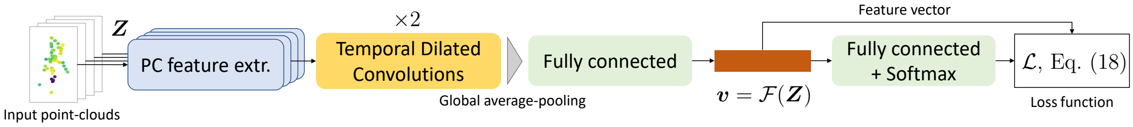


Figure 3.6: Block diagram of the NN feature extractor.

3.4.8 Extraction of Feature Vectors from mmWave Point-Clouds

To extract the gait features of the subjects, the NN proposed in [6], which was originally developed for person identification, is here adapted. The network uses a point-cloud feature extraction block inspired by PointNet [33], and followed by temporal dilated convolutions [34] to capture features related to the movement evolution in time. The proposed NN takes as input a radar point-cloud sequence, denoted by $\mathbf{Z}$, and outputs the corresponding feature vector $\mathbf{v} = \mathcal{F}(\mathbf{Z})$. Fig. 3.6 shows the block diagram of the NN. First, the network is expanded with respect to [6], using augmented point-cloud feature extraction blocks composed of 3 shared MLPs of size 98 and 2 MLPs of size 196, yielding point cloud features of size 196×1. Then, 2 temporal convolution blocks are used, containing 3, 3×3, convolutional layers each, with $(32, 64, 128)$ and $(256, 128, 32)$ filters, respectively, for the two blocks, and dilation rates of $1, 2, 4$ for the 3 layers in each block. Then, after applying the same global average pooling operation of [6], a fully connected layer [53] is introduced before the classification output, which produces a vector $\tilde{\mathbf{v}}$ of dimension 32. The final feature vector is obtained using L_2-normalization on $\tilde{\mathbf{v}}$, i.e., $\mathbf{v} = \tilde{\mathbf{v}}/||\tilde{\mathbf{v}}||_2$. A summary of the NN layers and their parameters is provided in Tab. 3.1.

Training

The NN is trained to produce representative feature vectors, $\mathbf{v}$, containing information on the way of walking of the subjects. This requires that the network generalizes well to subjects *not seen* at training time, as the performance of the re-identification mechanism strongly depends on the quality of the extracted features. To this end, in this chapter the NN is trained using a weighted combination of the *cross-entropy loss* [53], denoted by $\mathcal{L}_{\rm ce}$, the *center loss* [91], $\mathcal{L}_{\rm cnt}$, and the *triplet loss* [92], $\mathcal{L}_{\rm tri}$.

The cross-entropy is the most widely used loss for classification purposes in deep learning, and here it is used to train the network to distinguish among the different subjects [53]. However, just training the NN on a classification problem does not lead to sufficiently discriminative features for the re-identification mechanism. The center loss is adopted to additionally force the feature representations belonging to the same class to be close in the feature space, in terms of Euclidean distance. Specifically, denoting by $\mathbf{c}_l$ the centroid of the feature vectors belonging to class l, the center loss is

$$\mathcal{L}_{\rm cen}(\mathbf{v}, l) = ||\mathbf{v} - \mathbf{c}_l||_2^2, \tag{3.16}$$

where the centroids are learned as part of the training process via the back-propagation algorithm [91].

The triplet loss is used to push apart the feature representations of inputs belonging to different classes. For this, triplets of input samples are selected from the training set, two of them from the same class, leading to feature vectors $\mathbf{v}_a$ and $\mathbf{v}_b$, and one belonging to a different class, leading to a third feature vector $\mathbf{v}_c$. For further details on the triplet selection process, see Section 3.2 of [92]. The triplet loss is written as

$$\mathcal{L}_{\text{tri}}(\mathbf{v}_a, \mathbf{v}_b, \mathbf{v}_c) = \max \left\{ ||\mathbf{v}_a - \mathbf{v}_b||_2^2 - ||\mathbf{v}_a - \mathbf{v}_c||_2^2 + \mu, 0 \right\}, \tag{3.17}$$

where μ is a margin hyperparameter, set to 1. Hence, the feature extractor is trained with the following total loss function

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \omega \mathcal{L}_{\text{cen}}, \tag{3.18}$$

where the parameter $\omega = 0.5$ weighs the relative importance of the center loss. In the implementation, a training dataset containing mmWave radar point-clouds from 16 subjects is used. It was collected in different indoor environments to increase the generalization capabilities of the NN. The optimization is carried out using Adam [53] with learning rate 10^{-4} and an L_2 regularization rate of 8×10^{-5} for 250 epochs, as summarized in Tab. 3.1. Hyperparameters tuning was carried out using a greedy search procedure, optimizing the value of the loss $\mathcal{L}$ on a validation set containing a randomly selected subset (20%) of the training data.

Feature extraction

At inference time, i.e., during the system operation, the NN is used to compute feature vectors that are representative of the subjects' gait. Specifically, 45 steps (3 seconds) long sequences of radar point-clouds are collected for each tracked subject. The point-cloud sequences are denote by $\mathbf{Z}$ in the following. The inner representation $\mathbf{v} = \mathcal{F}(\mathbf{Z})$, after L_2-normalization, is used as the feature vector for the following re-identification mechanism.

3.4.9 Weighted Extreme Learning Machine (WELM)

The WELM [86] is a particular kind of single-layer feedforward neural network in which the weights of the hidden nodes are chosen randomly, while the parameters of the output layer are computed analytically. Consider an n_{cls}-class classification problem, a training set $\mathcal{V} = \cup_{n=1}^{n_{\text{cls}}} \mathcal{V}_n$ of input *feature vectors* $\mathbf{v}$ (see Section 3.4.8), each with an associated one-hot encoded label $\mathbf{y} \in \{0, 1\}^{n_{\text{cls}}}$, where $\mathcal{V}_n$ is the set containing the vectors from class $n = 1, \dots, n_{\text{cls}}$. For any $\mathbf{v} \in \mathcal{V}$, the WELM computes the matrix of hidden feature vectors $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times L}$, with rows $\mathbf{h}(\mathbf{v})$, where L is the number of WELM hidden units and $\mathbf{h}(\cdot)$ is a non-linear activation function. milliTRACE-IR uses $\mathbf{h}(\mathbf{v}) = \text{ReLU}(\mathbf{W}\mathbf{v} + \mathbf{b})$ where ReLU is the rectified linear unit [53] ($\text{ReLU}(x) = \max(x, 0)$) and $\mathbf{W}, \mathbf{b}$ are the weights and biases of the Extreme Learning Machines (ELM) hidden layer, respectively. The elements of $\mathbf{W}$ and $\mathbf{b}$ are here generated from $\mathcal{N}(0, 0.1)$. The WELM learning process amounts to computing, for each class n, the optimal values of an output weight vector $\boldsymbol{\beta}_n$ that minimizes the *weighted* LS L_2-regularized quadratic cost function $||\mathbf{H}\boldsymbol{\beta}_n - y_n||_{\boldsymbol{\Omega}}^2 + \lambda ||\boldsymbol{\beta}_n||_2^2$, where λ is a regularization parameter and $\boldsymbol{\Omega}$ is a diagonal weighting matrix used to boost the importance of those samples belonging to under-represented classes. This compensates for the tendency of the standard ELM to favor over-represented classes at inference time [86]. In the analyzed scenario, the individuals move freely in the environment across different rooms, so the number of feature vectors collected from each of them is not only unknown in advance, but highly variable. Hence, the training set usually contains unbalanced classes, and milliTRACE-IR uses

$$\Omega_{i,i} = 1/|\mathcal{V}_{n_i}|,\ i = 1, \dots, |\mathcal{V}|, \tag{3.19}$$

where $n_i = \text{argmax}_n(\mathbf{y}_i)_n$ denotes the class of the i-th vector. Stacking all the $\boldsymbol{\beta}_n$ into a single matrix $\mathbf{B} \in \mathbb{R}^{L \times n_{\text{cls}}}$ and the labels into matrix $\mathbf{Y} \in \{0, 1\}^{|\mathcal{V}| \times n_{\text{cls}}}$, the WELM output weights $\mathbf{B}$ can be computed in closed-form using one of the following equivalent expressions

$$\mathbf{B} = \mathbf{H}^T \left(\lambda \mathbf{I} + \boldsymbol{\Omega} \mathbf{H} \mathbf{H}^T\right)^{-1} \boldsymbol{\Omega} \mathbf{Y}, \text{ or} \tag{3.20}$$

$$\mathbf{B} = \left(\lambda \mathbf{I} + \mathbf{H}^T \boldsymbol{\Omega} \mathbf{H}\right)^{-1} \mathbf{H}^T \boldsymbol{\Omega} \mathbf{Y}. \tag{3.21}$$

Due to the dimension of the matrix to be inverted, if $|\mathcal{V}|> L$, it is more convenient to use Eq. (3.21), while if $|\mathcal{V}|\leq L$ Eq. (3.20) has to be preferred. The output classification for a vector $\mathbf{v}$ is then computed as $\text{argmax}_i \left(\mathbf{h}(\mathbf{v})^T \mathbf{B}\right)_i$, where $\mathbf{h}(\mathbf{v})^T \mathbf{B}$ is a vector of WELM scores for each class.

3.4.10 WELM based Person Re-Identification

To enable person re-identification based on the feature vectors $\mathbf{v}$ extracted by the NN, milliTRACE-IR uses the WELM multiclass classifier of Section 3.4.9, which is *trained at runtime* only when the system has to re-identify a previously seen subject. This is done by sequentially collecting feature vectors from all the subjects seen by the system at operation time, and storing them into the training set $\mathcal{V}$.

Note that, although an online sequential version of the ELM training process has been proposed in [93], the WELM is trained every time a person has to be re-identified using a batch implementation and including in the training set $\mathcal{V}$ all the subjects seen up to the current time-step k. This is because in the online training procedure of [93] the number of classes has to be *fixed* in advance, while in the considered setup the number of subjects seen by the system may change in time and the Re-Id procedure must be flexible to the addition of new individuals to the training set $\mathcal{V}$. The WELM training and re-identification phases are detailed next and in Alg. 3.2.

Training

The training process is performed at runtime as explained in Section 3.4.9, using $L = 1,024$ and $\lambda = 0.1$. During the normal system operation, the feature vectors obtained from each track are continuously added to set $\mathcal{V}$, storing the corresponding one-hot encoded vectors containing the subjects' identities into matrix $\mathbf{Y}$. To reduce the computational burden, the feature extraction step is executed every 5 time-steps. This is reasonable, as the input sequences to the NN contain 45 time-steps overall and extracting the features at every time-step would lead to highly correlated, and

Algorithm 3.2 Re-identification mechanism at time k.

```
Input: Training set $\mathcal{V}$, track to be re-identified $t^{\text{id}}$.
Output: Re-id label of $t^{\text{id}}$.
 1: $\mathbf{H} \leftarrow \left[\mathbf{h}^T(\mathbf{v}), \forall\, \mathbf{v} \in \mathcal{V}\right]$
 2: $\mathbf{Y} \leftarrow$ labels of $\mathcal{V}$
 3: $\boldsymbol{\Omega} \leftarrow$ Eq. (3.19)
 4: $\mathbf{B} \leftarrow$ Eq. (3.21) or Eq. (3.20) depending on $|\mathcal{V}| \lesseqgtr L$
 5: $\boldsymbol{\xi}_0 \leftarrow \mathbf{0}$
 6: for $j = 1, \dots, W$
 7:     $\mathbf{v}_j^{\text{id}} \leftarrow \mathcal{F}(\mathbf{Z}_j)$
 8:     $\boldsymbol{\xi}_j \leftarrow \left[\mathbf{h}^T(\mathbf{v}_j^{\text{id}})\mathbf{B} + j\boldsymbol{\xi}_{j-1}\right]/(j+1)$
 9: end for
10: label $\leftarrow \arg\max_i(\boldsymbol{\xi}_W)_i$
```

therefore less informative feature vectors, in addition to entailing a higher computation cost. At time-step k, if a subject has to be re-identified, the training procedure of Section 3.4.9 is executed (lines $1-4$): the WELM feature vectors $\mathbf{H}$ are computed by applying the activation function $\mathbf{h}(\cdot)$ to each training vector and the weight matrix $\boldsymbol{\Omega}$ is obtained from Eq. (3.19) (lines $1-3$). The WELM output matrix $\mathbf{B}$, is computed using Eq. (3.20) or Eq. (3.21) depending on $|\mathcal{V}|$ (line 4).

Re-identification

The Re-Id procedure is used to recognize subjects that have been seen by the system and associate them with their temperature measurement and their past movement history in the monitored area. Denoting by t^{id} the track to be re-identified, the trained WELM processes the NN features of this user, $\mathbf{v}^{\text{id}}$, as follows: $\mathbf{h}(\mathbf{v}^{\text{id}})^T\mathbf{B}$. Due to the high variability of human movement, rather than considering a single feature vector, milliTRACE-IR computes the cumulative average WELM scores over a time window of length W, where the average score at time $j = 1, \dots, W$ is referred to as $\boldsymbol{\xi}_j$ (lines $6-9$). The identity label corresponds to the index of the largest element of $\boldsymbol{\xi}_W$ (line 10).

3.5 Experimental Results

In this section, the experimental results obtained by testing the system in different indoor environments are presented.

3.5.1 Implementation

Hardware. milliTRACE-IR has been implemented on an NVIDIA Jetson TX2 edge computing device[‡], with 8 GB of RAM and a NVIDIA Pascal GPU. The Jetson TX2 has been connected via Universal Serial Bus (USB) to a Texas Instruments IWR1843BOOST mmWave radar[§], operating

[‡] https://developer.nvidia.com/embedded/jetson-tx2

[§] https://www.ti.com/tool/IWR1843BOOST

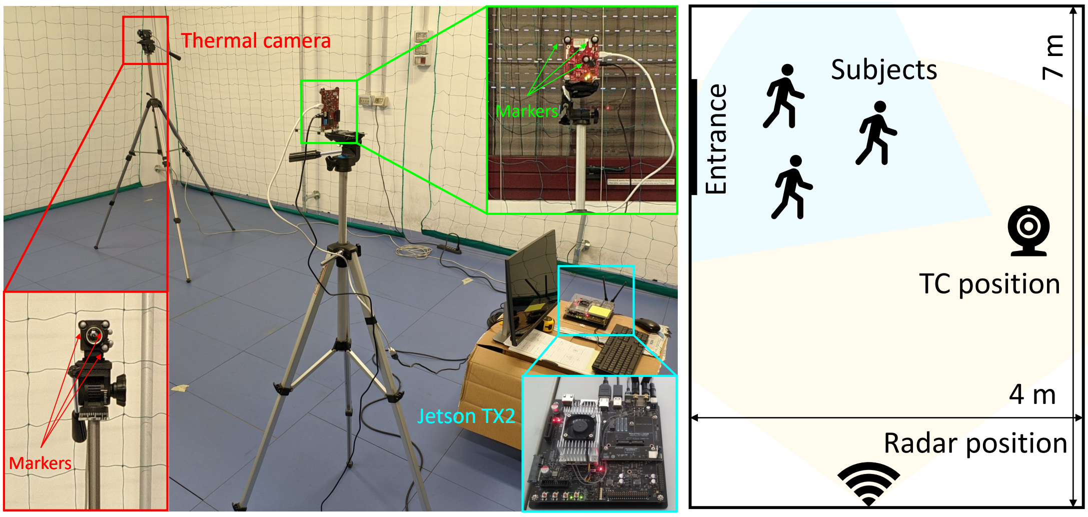


Figure 3.7: Experimental setup for the data association.

in the $77 - 81$ GHz band, and via Ethernet to a FLIR A65 thermal camera[¶], as shown in Fig. 3.7. The experiments have been performed in real-time at a frame rate of $1/\Delta = 15$ Hz.

The radar device operates in FMCW mode, using a chirp bandwidth $B = 3.07$ GHz, which leads to a range resolution of $c/2B = 4.88$ cm, and 64 chirps per sequence, obtaining a maximum measurable velocity of 4.77 m/s and velocity resolution of 14.92 cm/s.

The thermal camera has a 640×512 Focal Plane Array (FPA), a spectral range of $[7.5, 13]$ μm, a temperature range of $[-25, 135]$°C, a measurement uncertainty of ±5°C, and a Noise Equivalent Temperature Difference (NETD) of 50 mK.

Software. The system has been developed in Python, using the NumPy, SciPy and OpenCV libraries for the implementation of the tracking phases (for radar and thermal camera) and the proposed data association (Section 3.4.6), clustering (Section 3.4.5) and re-identification (Section 3.4.10) algorithms. Tensorflow and Keras libraries have been used to implement the feature extraction NN (Section 3.4.8). The pre-trained face detector for the thermal images (Section 3.4.2) has been taken from the open-source YOLOFace[‖] implementation.

3.5.2 TC and Radar Tracks Association

To assess the performance of the radar-TC track association method, experimental tests were conducted in a 7×4 m research laboratory. A motion tracking system including 10 cameras was used to gather Ground Truth (GT) data about the locations of the subjects, by placing markers atop their heads. This camera based tracking system provides 3D localization with millimiter-level precision, for all markers, at a rate of 100 Hz. The radar and the TC were placed as shown in Fig. 3.7. 5 measurement sequences with 2 subjects and 9 sequences with 3 subjects, all freely

[¶] https://www.flir.it/products/a65/

[‖] https://github.com/sthanhng/yoloface

	With $\rho(K)$		Without $\rho(K)$	
	Pr [%]	Rec [%]	Pr [%]	Rec [%]
$A_x + A_d$	**97.3**	**97.3**	91.9	91.9
A_x only	91.9	89.2	89.7	94.6
A_d only	92.1	94.6	86.8	89.2

Table 3.2: Impact of the components of the cost function. Row labels A_x, A_d, and $A_x + A_d$ indicate, respectively, that only costs A_x, A_d or the sum of the two were used in the evaluation. Label "With $\rho(K)$" indicate that the corrective term, $\rho(K)$, was used, while label "Without $\rho(K)$" means $\rho(K) = 1$.

entering the room, were collected. The roto-translation matrix $\mathbf{\Phi}$ was estimated using a set of markers applied to the devices, while the TC intrinsic matrix $\mathbf{\Psi}$ (see Section 3.4.6) and the radial distortion coefficients were obtained through the Zhang's method [94], using a sun-heated checkerboard pattern.

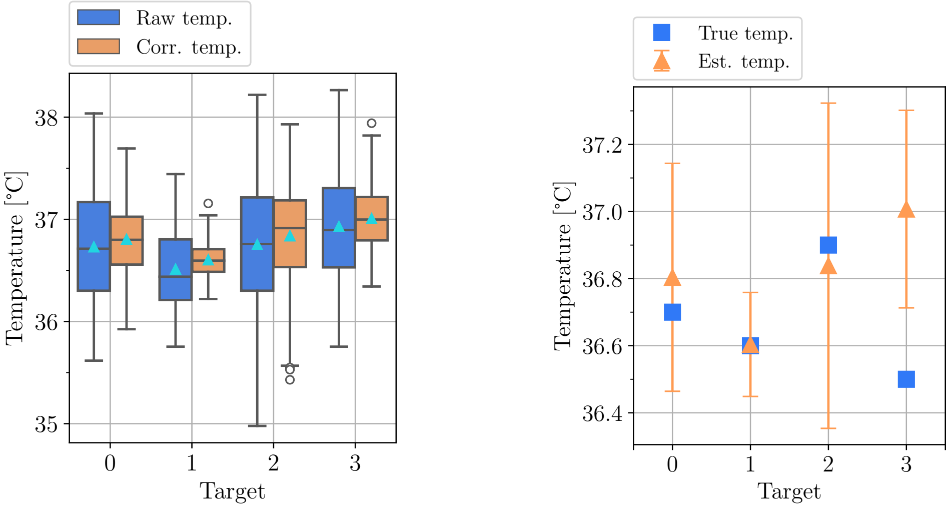
An *association* is defined as a specific pairing $i \leftrightarrow j$ of a track i from the radar with a track j from the TC, and a *correct association* as an association for which the two tracks correspond to the same subject. Given a set of tracks, the set of all the correct associations performed by the algorithm is denoted by $\mathcal{A}_{\text{TP}}$ (*true positives*), the set of all the associations performed by the algorithm as $\mathcal{A}_{\text{P}}$ (*positives*), and the set of all the associations that the algorithm should have performed, based on the GT, as $\mathcal{A}_{\text{R}}$ (*relevant*).

To quantify the association performance of the system, define the *precision*, $\text{Pr} = |\mathcal{A}_{\text{TP}}|/|\mathcal{A}_{\text{P}}|$, and the *recall*, $\text{Rec} = |\mathcal{A}_{\text{TP}}|/|\mathcal{A}_{\text{R}}|$. Using these metrics, the proposed track association method is evaluated by assessing the contribution of each cost component in $A(i, j)$ (see Eq. (3.11)). The results are reported in Tab. 3.2, where the row labels A_x, A_d, and $A_x + A_d$ indicate the cost function used. The table also shows the impact of adding the correction coefficient $\rho(K)$ (see Eq. (3.10)): for the case "Without $\rho(K)$", $\rho(K)$ is set to 1.

As shown, the proposed track association method reliably associates the radar and TC tracks, reaching precision and recall both higher than 97%. The joint use of A_x, A_d and $\rho(K)$ leads to improvements of up to 11% and 8% for the precision and recall metrics, respectively.

3.5.3 Temperature Screening

Remarkably, the proposed temperature screening method does not require people to stand in front of the TC sensor, but estimates their temperature as they move within the FoV of the TC. In order for the method to return accurate temperature measurements, the subject' frontal face should be captured by the TC for a minimum time duration. For this reason, it is advisable to place the TC near a point of passage, e.g., in proximity of an entrance. The temperature screening method was tested on $4 - 7$ sequences of ~ 10 s each were collected from 4 different individuals moving within 3.5 m from the TC. Each subject was tested at a different time of the day, to gauge the effects of the changing (thermal) environmental conditions, and of a possible concept drift (e.g., heating) of the TC after a long period of operation. Furthermore, as explained in Section 3.4.3, a



(a) Comparison between with (*Corr. temp.*) and without (*Raw temp.*) distance-based correction. The triangles show the mean values.

(b) Comparison between the estimated temperatures and the true temperatures. The error bars represent the standard deviations.

Figure 3.8: Results of the temperature screening.

	Mean [°C]	± std [°C]	True temp. [°C]	Error [°C]
Target 0	36.8	0.340	36.7	0.104
Target 1	36.6	0.155	36.6	0.004
Target 2	36.8	**0.485**	36.9	−0.062
Target 3	37.0	0.294	36.5	**0.507**

Table 3.3: Results of the temperature estimation and comparison with respect to the true values for the 4 targets. The worst cases are highlighted.

linear function $\alpha(\cdot)$ was fit to compensate for the influence of the distance on the measures.

To evaluate the benefit brought by the correction based on the targets' distance, in Fig. 3.8a, the results obtained with (*Corr. temp.*) and without (*Raw temp.*) the correction are compared. Since the TC is intrinsically subject to a bias, to facilitate the comparison of the measures, in the *Raw temp.* case only this bias is corrected, assuming a constant target distance of 2 m and multiplying each measured temperature by $\alpha(2) = a_0 + 2a_1$. The full method (*Corr. temp.*), instead, uses the rescaled average estimate, as per Eq. (3.15). The box-plot shows that the range of the corrected temperatures is significantly reduced (for these experiments, the true temperature is constant), demonstrating the efficacy of the proposed correction plus averaging approach. As an illustrative example, Fig. 3.9 shows the impact of the distance-based correction on data measurements from a subject moving in front of the TC.

Fig. 3.8b compares the temperature estimates from milliTRACE-IR and the true temperatures measured with a contact thermometer. The numerical results are reported in Tab. 3.3, where

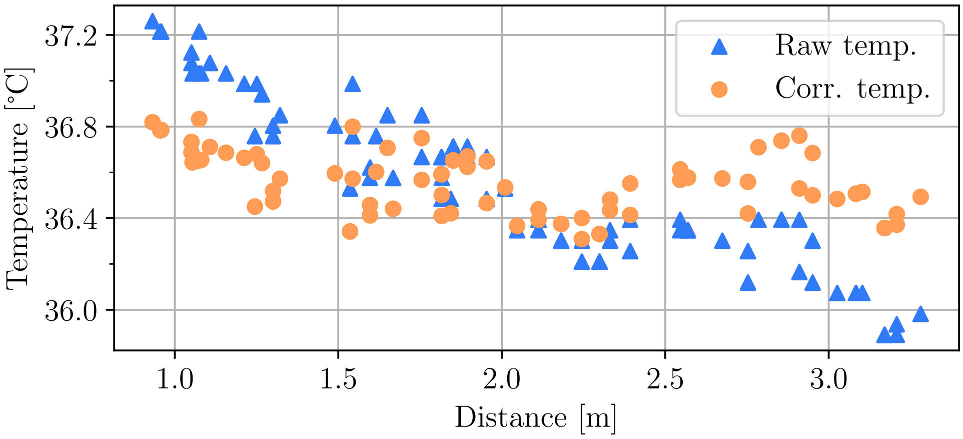


Figure 3.9: Temperature measurements from a subject moving in front of the TC with (*Corr. temp.*) and without (*Raw temp.*) distance-based correction.

the worst cases are reported in bold fonts. Mean temperatures are estimated with a maximum standard deviation from the mean smaller than 0.5 °C and a maximum absolute error with respect to the true temperature of about 0.5 °C. Note that only one of the subjects in Fig. 3.8b exhibits this maximum error (subject 3), while the absolute error for the others remains within 0.1 °C. These errors descend from the fact that the environmental conditions and the heating of the thermal camera affect the measurements in an unpredictable way, modifying the *bias* of the fitting function. Notwithstanding, the thermal screening capability of milliTRACE-IR is significantly better than that of existing approaches, see Section 3.5.7. Also, some improvements are possible by, e.g., applying a correction based on an external reference, such as a piece of material instrumented with a contact thermometer and located within the field of view of the TC, or monitoring the statistics of the people's temperature (mean μ and standard deviation σ) to detect anomalous samples within such empirical distribution. For instance, an alarm could be raised for those subjects whose temperature is greater than $\mu + c \times \sigma$, for a user-defined threshold c. This would allow the system to continuously and autonomously adapt to different operating conditions.

3.5.4 Positioning and Social Distance Monitoring

To evaluate the performance of the radar tracking system in estimating the position of the targets and the inter-subject's distance, tests were conducted in the 7×4 m research laboratory described in Section 3.5.2. A total of 7 sequences of duration $10 - 15$ s were collected, each with 3 subjects moving freely in the room, along with their GT locations obtained from the motion tracking system. The Root Mean Squared Error (RMSE) between the mmWave radar estimated locations and the GT is used as a performance metric. Moreover, the inter-subject distances were measured, considering all the possible combinations of the three subjects and leading to a total of 21 inter-subject distances across all the recorded sequences.

The Cumulative Distribution Function (CDF) of the absolute error between the ground truth

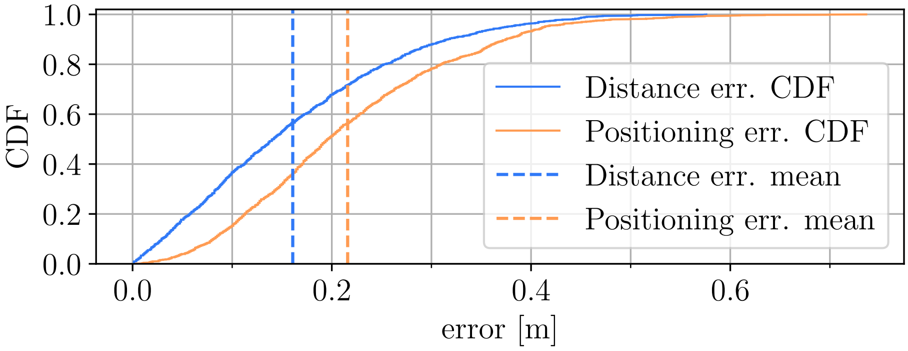


Figure 3.10: CDF of the absolute error between the true (ground truth) and the estimated subject's position / inter-subject's distance, as measured by the radar tracking system. The dashed lines denote the mean error.

	Mean [m]	± std [m]	Frames	Time [s]
Position RMSE	0.216	0.115	1448	97
Subj. distance RMSE	0.161	0.112	1153	77

Table 3.4: RMSE of the subject's position and of the inter-subject's distance estimated by the radar sensor, computed against the GT.

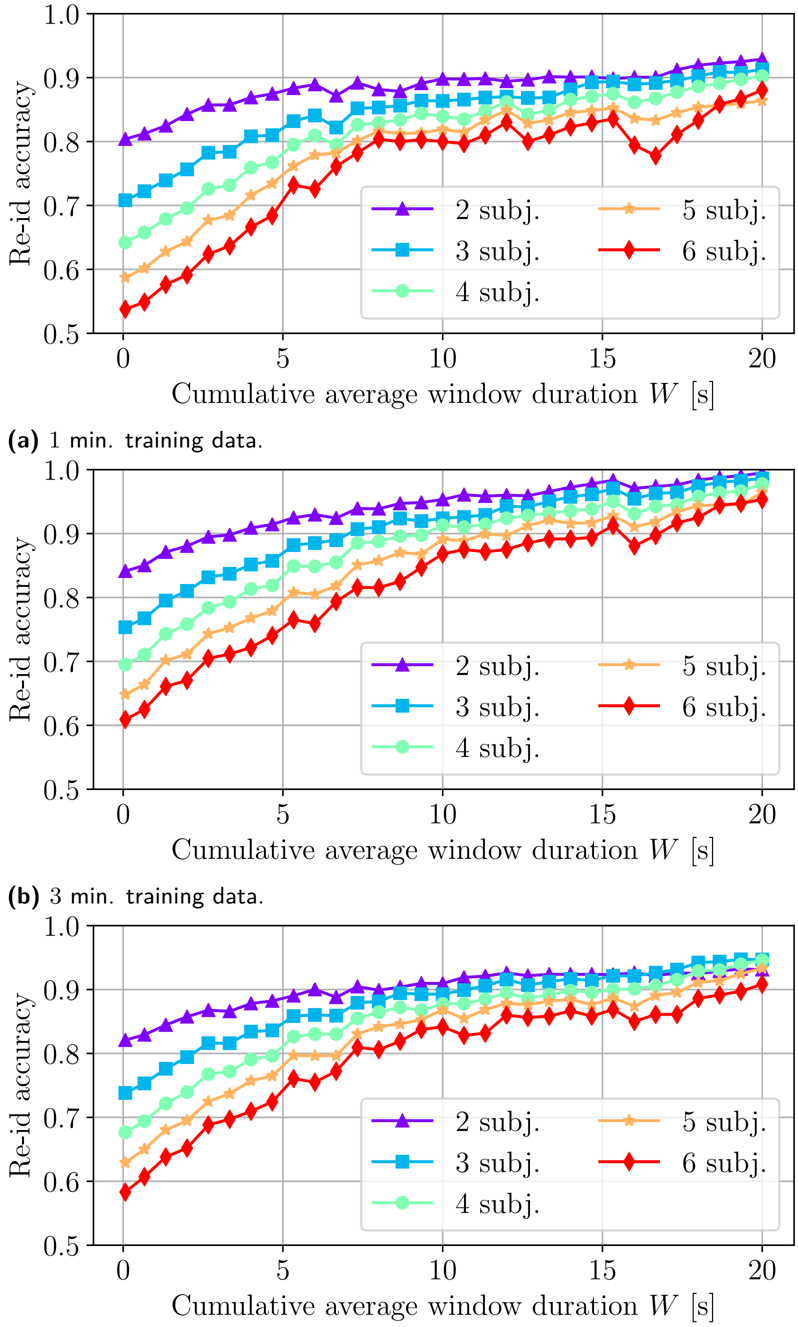
	milliTRACE-IR		**DBSCAN**	
	r_{cl} [%]	corr. tracked	r_{cl} [%]	corr. tracked
2 sub. parallel	90.7	✓	46.5	×
2 sub. crossing	87.9	✓	59.6	×
2 sub. close	89.9	✓	69.7	✓
3 sub. parallel	92.3	✓	65.3	✓
3 sub. crossing	83.7	✓	73.5	×

Table 3.5: Ratio r_{cl} between the number of frames in which the different subjects are correctly separated and the total number of frames, using the proposed method and DBSCAN. Symbols "✓" and "×" denote success and failure of the tracking step, respectively.

and the estimated subject's position/inter-subject distance, as measured by the radar tracking system, is shown in Fig. 3.10, along with the corresponding mean values. The numerical results are provided in Tab. 3.4. The radar system achieves an absolute *positioning* error within 0.3 m in 80% of the cases. For the inter-subject *distance*, the error remains within 0.25 m in 80% of the cases.

3.5.5 Effectiveness of the Improved Clustering Technique

To evaluate the improvement brought by the proposed clustering method over the standard DBSCAN, both algorithms were tested on specific measurement sequences with subjects moving



(a) 1 min. training data.

(b) 3 min. training data.

(c) Imbalanced training data.

Figure 3.11: Re-identification accuracy results. In (a) and (b) the re-identification algorithm is used with 1 and 3 minutes of training data per subject, respectively. In (c), 1 minute of training data was used for a randomly selected subset containing half of the subjects, while 4 minutes were used for the remaining half.

within 1 m from one another. To quantify the clustering performance, the correct clustering ratio, r_{cl}, is used. This metric represents the fraction of frames in which the clusters belonging to the different subjects are correctly separated. The results of this evaluation are summarized in Tab. 3.5. The evaluation is conducted on sequences with 2 and 3 individuals *(i)* walking along parallel paths with the same velocity and at a distance between 0.5 m and 0.8 m (*parallel*), *(ii)* walking along crossing paths, with subjects coming as close as 0.2 m from one another (*crossing*) and *(iii)* staying still and moving arms at an inter-subject distance of approximately 0.8 m (*close*). The proposed clustering algorithm led to a large improvement (up to 44 %) in terms of r_{cl} metric with respect to DBSCAN. In addition, for 3 of the 5 test sequences, DBSCAN led to failures in the tracking process, either merging the tracks of different subjects, or failing to detect some of them, while milliTRACE-IR correctly tracked all the subjects in all cases.

3.5.6 Person Re-Identification

The proposed WELM based Re-Id algorithm was evaluated on a set of mmWave radar measurements from 6 individuals who were *not* included among the 16 subjects used to train the feature extraction NN. The tests were conducted in a 12 × 3 m research lab, with furniture that made the evaluation challenging. The training data contains 4 minutes of measurements ($3,600$ radar frames) while over 1 minute of measurements per subject ($1,000$ frames) was used as test data. In both the training and the test data, the individuals walked freely in the room. The radar position was changed for each test to gauge the impact of varying the radar point-of-view.

Re-Id accuracy. The Re-Id accuracy as a function of W (see Alg. 3.2) is shown in Fig. 3.11a and Fig. 3.11b. The curves of these plots are obtained averaging the results of 20 different WELM initializations, and all the possible combinations of the considered number of subjects (from 2 to 6) over the 6 total individuals. As expected, the Re-Id performance increases with an increasing inference time (larger W) and with the length of the training sequences: the accuracy gain is about 10% by going from 1-minute (Fig. 3.11a) to 3 minutes (Fig. 3.11b) long training sequences. Also, milliTRACE-IR reaches high Re-Id accuracy using $W \geq 15$ s and the detrimental effect of increasing number of subjects to be classified is greatly reduced using larger values of W, as accumulating the WELM scores over longer time windows increases the robustness of the WELM decision. Overall, the accuracy of the proposed method is higher than 95% in all cases, only using 3 minutes of training data per subject and $W = 20$ s, which are reasonable in practice. The worst-case (3 minutes of training data for 6 subjects) WELM training time, on the ARM Cortex-A57 processor of the Jetson TX2 device, took 2.98 ± 0.015 s.

Impact of imbalanced training data. As shown in Fig. 3.11c, the effect of imbalanced training data is successfully mitigated by the sample weighting strategy of Eq. (3.19). In this evaluation, the WELM was trained with 1 minute of data for a randomly selected subset containing half of the subjects and 4 minutes for the remaining half.

Improvement over a baseline. Tab. 3.6 compares the WELM to a baseline classification method widely used in camera-based person Re-Id [16] that, unlike milliTRACE-IR, does not learn a similarity score based on the actual distribution of the feature vectors at operation time.

	WELM			**Cos. sim. baseline**		
	1 min.	4 min.	imb.	1 min.	4 min.	imb.
$W = 0$ s	53.8	60.9	58.3	44.6	49.5	51.6
$W = 10$ s	80.0	86.8	84.2	63.9	77.7	80.6
$W = 20$ s	88.6	95.3	90.8	72.2	88.8	86.9

Table 3.6: Re-Id accuracies obtained by the WELM and the CS baseline on 6 subjects using 1 and 4 minutes balanced training sets, and an imbalanced training set. The cumulative average window W is set to 0 s (a single test feature vector is used), 10 s or 20 s.

	milliTRACE-IR	**Ulrich** [81]	**Savazzi** [68]
Positioning range RMSE [m]	0.19	×	0.45
Interpersonal dist. RMSE [m]	0.17	×	0.5*
Positioning angle RMSE [°]	3.1	×	7.0
Thermal screening RMSE [°C]	0.13	×	0.45
Thermal screening range [m]	3.5	×	1.1
Per-frame assoc. Pr [%]	98.6	25.2	n.a.
Per-frame assoc. Rec [%]	98.5	78.4	n.a.
Re-identification acc. [%]	≈ 90	×	×

Table 3.7: Comparison of milliTRACE-IR with the works from Ülrich et al. [81] and Savazzi et al. [68]. Symbols "×" and "n.a." denote, respectively, that the task is not tackled or that there is no available result for the considered quantity in the original papers. The symbol "*" is used to highlight that the value is not an RMSE value but the minimum interpersonal distance threshold considered in [68].

The baseline algorithm collects the training feature vectors along with the corresponding labels and computes the *centroid* of each class m in the NN feature space, denoted by $\mathbf{c}_m$. To re-identify a subject, the cosine similarity between his/her feature vectors, $\mathbf{v}$, and the centroid of each class m is computed, obtaining a similarity score $s_m = \mathbf{c}_m^T\mathbf{v}/(||\mathbf{c}_m||_2 \times ||\mathbf{v}||_2)$, and the classification is performed taking $\text{argmax}_m\, s_m$. The WELM outperforms the baseline scheme in all the tests, see Tab. 3.6. The performance gap is significant for little training data (up to 16% improvement), small windows and imbalanced training sets.

3.5.7 Comparison with existing approaches

In this section a comparison between milliTRACE-IR and available methods from the literature is provided. To the best of the authors' knowledge, only two works exploit both mmWave radars and thermal cameras to perform human sensing and/or temperature screening, namely, the works from Ülrich et al. [81] and Savazzi et al. [68]. Since none of the two tackles all the points that milliTRACE-IR addresses, they are here considered, separately, to compare different aspects. The data association strategy is compared with that proposed in [81], while [68] is used to compare the positioning, distance monitoring, and temperature screening parts. In Tab. 3.7, symbols "×"

and “n.a.” denote, respectively, that the task is not tackled or that no specific result is provided in the corresponding work.

Data association. In [81] (Ülrich et al.), people are detected in thermal images by applying the Viola and Jones algorithm [95] to detect the upper bodies of the subjects in the environment. The distance between the TC and each subject is roughly retrieved from the dimension of the bounding box enclosing the upper body of each person, similarly to what milliTRACE-IR does with faces. The TC detections are then associated, on a frame basis, with range measurements obtained with a mmWave radar by minimizing a Gaussian-shaped association cost. This cost provides an estimate of the probability that the corresponding association is correct, based on the difference between the distance estimates by the TC and by the mmWave radar. This data association method has been implemented and tested on the dataset of Section 3.5.2, comparing it to the data association strategy of milliTRACE-IR. For a fair comparison, the YOLOv3 detector has been used in place of the Viola and Jones algorithm, as, besides providing superior performance, it is the same detector used by milliTRACE-IR. This guarantees that any difference in the data association results is only due to the data association strategy. At every time frame, each bounding box has been associated with the radar detection yielding the highest association probability, which corresponds to the smallest difference in the two distance estimates. The main differences between the approach in [81] and that of milliTRACE-IR are that, in [81]: *(i)* the association is per-frame and not per-track, *(ii)* the estimated distance from the TC is the only feature considered for the association, and *(iii)* the Hungarian algorithm is not used, so different bounding boxes can be, erroneously, associated with the same radar detection. Numerical results for the precision (“Pr”) and recall (“Rec”) metrics are presented in Tab. 3.7. Since the association technique of [81] performs a per-frame association, the table shows the per-frame performance of milliTRACE-IR, computed by counting the number of frames that are correctly classified using milliTRACE-IR’s per-track association algorithm. From these results, it can be seen that milliTRACE-IR performs notably better in associating mmWave radar with TC human detections. The largest improvement is brought by the combination of milliTRACE-IR *per-track* association paradigm with the Hungarian algorithm, which effectively filters out ghost tracks and spurious detections which often occur in real world scenarios, significantly boosting the robustness of the scheme.

Positioning, distancing, and temperature screening. In [68] (Savazzi et al.), people localization, interpersonal distance monitoring, and temperature screening are addressed using thermopiles and mmWave radars. Since in [68] the data association strategy is not disclosed, a comparison is here provided only for the previously mentioned tasks. In the paper, positioning performance is evaluated in terms of range (radial distance) and angular RMSEs. Numerical values for these metrics are given in Tab. 3.7 considering the dataset of Section 3.5.4 for milliTRACE-IR and the (average) values from Tab. II of [68] for their algorithm.

In the same work, interpersonal distance monitoring is obtained by dividing the monitored area into a regular grid, whose cells have a side length of 0.5 m. The system is able to distinguish subjects occupying adjacent cells, which are considered to be violating the minimum interpersonal distance of 1 m, thus raising an alarm. For this reason, the resolution of the method of [68] is 0.5 m in the best case (a lower bound for the interpersonal distance estimation error). In Tab. 3.7, this value is reported alongside the RMSE of milliTRACE-IR in measuring interpersonal distances, marking the former with a “*” symbol, to highlight that it is not an RMSE.

Thermal screening performance comparisons are also presented in Tab. 3.7, where “Thermal screening range [m]” refers to the maximum distance at which the tests were carried out. milliTRACE-IR performs better than [68] in all the considered tasks, showing a larger monitoring range and more accurate body temperature estimates. In addition, milliTRACE-IR combines these monitoring capabilities with a robust data association strategy and with the capability to re-identify subjects when moving through different areas.

3.6 Concluding Remarks

This chapter presents the design and implementation of milliTRACE-IR, the first system combining high resolution mmWave radar devices and infrared cameras to perform non-invasive joint temperature screening and contact tracing in indoor spaces. This system uses thermal cameras to infer the temperature of the subjects, achieving measurement errors within 0.5 °C, and mmWave radars to infer their spatial coordinates, by successfully locating and tracking subjects that are as close as 0.2 m apart. This is possible thanks to improvements along several lines, such as the association of the thermal camera and radar tracks from the same subject, along with a novel clustering algorithm combining density-based and Gaussian mixture methods to separate the radar reflections coming from different subjects as they move close to one another. Moreover, milliTRACE-IR performs contact tracing: a person with high body temperature is reliably detected by the thermal camera sensor and subsequently traced across a large indoor area in a non-invasive way by the radars. When entering a new room, this subject is re-identified among several other individuals with high accuracy (95%), by computing gait-related features from the radar reflections through a deep neural network and using a weighted extreme learning machine as the final re-identification tool.

4

Retrofitting IEEE 802.11ay Access Points for Indoor Human Detection and Sensing

4.1 Introduction

In this chapter we move on to the integration of mmWave radar sensing techniques in communication systems. Specifically, we propose RAPID, an ISAC platform that performs radar-like sensing of human movement based on the next generation IEEE 802.11ay Wi-Fi standard in the 60 GHz band. Previous works based on the precursor standard IEEE 802.11ad exploit the CIR estimation procedure for localizing people [28], [96], but they are not fully compliant with the communication packet structure specified by the standard and cannot match the sensing accuracy of radars, as no μD information is captured. In contrast, RAPID works without modifying the packet structure by leveraging the *in-packet* beam training and beam tracking features of IEEE 802.11ay. This leads to very low implementation and deployment cost, and allows for a highly accurate extraction of human movement information from the radio signals. IEEE 802.11ay uses highly directional antennas for communication. It specifies efficient *in-packet* beam training and tracking procedures [25], based on training (TRN) fields consisting of repetitions of complementary Golay sequences [26]. These fields are transmitted with different beam patterns, which allow determining which of the beam patterns is best for communication. By exploiting beam training packets, RAPID can accurately localize multiple human subjects within the same indoor space. Then, the μD signature associated with the movement of each subject is extracted by relying on the TRN units embedded in the data packets used for beam tracking, analyzing the phase differences of the CIR across subsequent packets that are reflected back by the environment. For such radar-like operation, RAPID IEEE 802.11ay APs have to enable their transmit and receive chains simultaneously, avoiding the problem of random phase offsets as transmitter and receiver share the same local oscillator. However, note that this does not require complex self-interference cancellation for full-duplex communication, since the receiver needs to only detect the highly robust Golay sequences of the TRN fields. Thanks to its unique design, RAPID is the first system that successfully extracts the μD from multiple subjects using standard Wi-Fi transmission technology, achieving radar-level accuracy. This is challenging, as it involves *(i)* striking a good balance between the packet transmission rate and the Doppler frequency resolution required to capture the μD of human movement, while *(ii)* ensuring sufficient phase coherence across adjacent packets. The obtained μD spectrograms are processed using a deep learning classifier to carry out continuous HAR and person identification.

Furthermore, we show that, thanks to the intrinsic superior ranging resolution of the mmWave spectrum, RAPID outperforms state of the art human sensing technology based on sub-6 GHz systems. Multiple moving subjects can be individually tracked, separating their signal reflections and, in turn, obtaining large improvements in terms of accuracy, robustness and generalization across environments and subjects. In addition, multiple RAPID-APs can be seamlessly integrated to boost detection and tracking performance. This also increases HAR and person identification accuracy by combining the information from different viewpoints. In this chapter, RAPID-APs were implemented using an FPGA-based SDR platform equipped with phased antenna arrays, which transmits IEEE 802.11ay-compliant packets and operates in a full-duplex fashion.

The main contributions of this chapter are:

1. We design and implement RAPID, a fully standard compliant multi-AP ISAC system that exploits IEEE 802.11ay TRN fields to achieve radar-like human sensing, including simultaneous multi-person tracking, HAR and person identification. RAPID reuses existing fields in the communication packets and avoids the need for a dedicated sensing infrastructure. In addition, it can also combine information from multiple APs for improved performance.

2. We propose a novel method to extract μD signatures of human movement from IEEE 802.11ay CIR estimates obtained from a sequence of packets, exploiting the Golay sequences specified in the standard. To the best of our knowledge, this is the first work to do so from 60 GHz communication waveforms.

3. We implement RAPID on a FPGA-based testbed including multiple IEEE 802.11ay-compliant APs which support full-duplex operation, so that each AP can listen to its own transmitted signal and act as a monostatic ISAC device.

4. We conduct an extensive measurement campaign in an indoor space to evaluate the proposed system and compare it to sub-6 GHz Wi-Fi systems. To this end, we build a unique dataset including simultaneous IEEE 802.11ay and IEEE 802.11ac CIR estimates. RAPID achieves continuous tracking of up to 5 concurrently moving subjects, with HAR accuracy of 94% and person identification accuracy of 90%. Moreover, it outperforms state of the art sub-6 GHz Wi-Fi sensing, showing superior accuracy and robustness to different environments and subjects.

The chapter is organized as follows. The related work is summarized in Section 4.2. RAPID is introduced in Section 4.3, presenting its constituent processing blocks. A brief summary of how IEEE 802.11ay can be used for environment sensing is given in Section 4.4, while in Section 4.5 the implementation of RAPID on FPGA hardware is discussed. A thorough performance analysis of RAPID on real measurements is presented in Section 4.6. Concluding remarks are presented in Section 4.7.

4.2 Related work

Sub-6 GHz sensing. Legacy Wi-Fi technologies such as IEEE 802.11n and IEEE 802.11ac, respectively working at 2.4 or 5 GHz, have been extensively used for human sensing, including activity/gesture recognition [20], [21], [97], [98], vital sign monitoring [99] and person identification [100].

Due to the rich multipath environment at lower frequencies, existing approaches have reached good accuracies by leveraging Orthogonal Frequency Division Multiplexing (OFDM) transmission and analyzing the CIR amplitude obtained at the different subcarriers, as done in [20]. The performance of such systems can be further improved by exploiting the phase components of the CIR [97], [98], but this entails using complex algorithms for the removal of random phase offsets.

Although there is a large body of work that exploits these technologies, they have two main drawbacks: either *(i)* they are effective for single-person scenarios, as the small available bandwidth only allows for coarse localization and tracking of the subjects, or *(ii)* they are highly sensitive to changes in the environment and hardly generalize to new scenarios (never seen at system calibration/training time), which can significantly worsen their performance. Addressing problem *(i)*, in [100], multi-person identification using IEEE 802.11n is achieved in a through-the-wall setting, but the subjects still need to be well separated in space (e.g., by at least 20° in azimuth angle at a distance of several meters). To mitigate the dependence on the environment, more elaborate deep learning and optimization approaches have been proposed in [98], [101], [102].

mmWave frequencies offer a natural solution to the above issues, by providing decimeter-level accuracy in distance measurements and high sensitivity to the μD effect, due to their small transmission wavelength. In addition, due to the sparsity of the mmWave channel, higher robustness to environmental changes is achieved.

mmWave radar. mmWave radars have been intensively studied in the past few years as an effective means to achieve fine-grained environment sensing [103]. Typical operating frequencies for these devices are the 60 or the 77 GHz bands. Centimeter-level accuracy in measuring distances is achieved thanks to the use of very large transmission bandwidths, up to 4 GHz, as dedicated radar devices are not constrained by communication requirements. Radars allow accurate HAR [3], [4] and have been used to perform person identification on small to medium-sized groups of people (up to a few tens), due to their very high resolution in obtaining the μD signatures of the subjects [6], [12]. In these works, the separation of the reflections from subjects concurrently moving in the environment is achieved through MIMO radars, which enable high angular resolution and allow tracking the users with errors below 0.2 m even in realistic scenarios where people walk and move freely [8]. However, these results are obtained within relatively small distances from the radar device, ranging from 4 [8] to $6-7$ m [12].

Despite the advanced sensing capabilities, mmWave radars entail high deployment costs to cover large indoor areas, even more considering their limited working range. For this reason, multi-radar networks to cover wider areas and avoid occlusions are seldom considered in the literature. Reusing existing mmWave communication links, as we do in this chapter, allows avoiding the costly deployment of additional hardware, while maintaining radar-like human sensing and detection performance.

802.11ad 60 GHz sensing. Commodity 60 GHz radios have been utilized for client device localization [104], people tracking [96], fine-grained human gesture recognition [105], [106], vital sign monitoring [107] and RF imaging [28]. All the works addressing human sensing are based on the IEEE 802.11ad standard and leverage the CIR estimation to obtain information about the environment. However, they typically do not address the problem of *joint* communication and sensing, which requires to reuse the packet structure specified by the communication standard. [105] and [106] address fine-grained hand gesture tracking. In [105], pulsed radar-like operations are performed to detect and track a human hand, reconstructing handwriting with centimeter-level accuracy. Notably, [106] performs similar processing using the IEEE 802.11ad CIR estimated by a mobile device for gesture classification. In [28], a commodity 60 GHz radio equipped with a 6×6 antenna array is used to obtain the silhouette of a person moving directly in front of the device. This is achieved with an angular super-resolution algorithm derived from MUSIC [108]. However, the device needs to be operated in a *radar mode* for transmission, which may not comply with the communication standard.

In [96], the estimated CIR amplitude is used along with receiver beamforming to localize and track multiple people, achieving a median localization error of 9.9 cm. This work does not exploit the phase of the CIR to extract the μD signature of the subjects, which is necessary to carry out HAR and person identification tasks. Moreover, the extension to the case of multiple APs is not considered.

Dedicated mmWave technology has been successfully exploited for human sensing, and new platforms are appearing regularly. However, to the best of our knowledge, RAPID is the first system that extracts radar-like μD signatures of human movement from IEEE 802.11ay APs, by retrofitting them with channel sensing and Doppler extraction capabilities. This is obtained by preserving the IEEE 802.11ay packet structure, obtaining a joint radar-communication platform that is fully standard compliant.

4.3 RAPID design

RAPID enables indoor human sensing in IEEE 802.11ay networks, by leveraging the network's *in-packet* beam training and beam tracking structure.

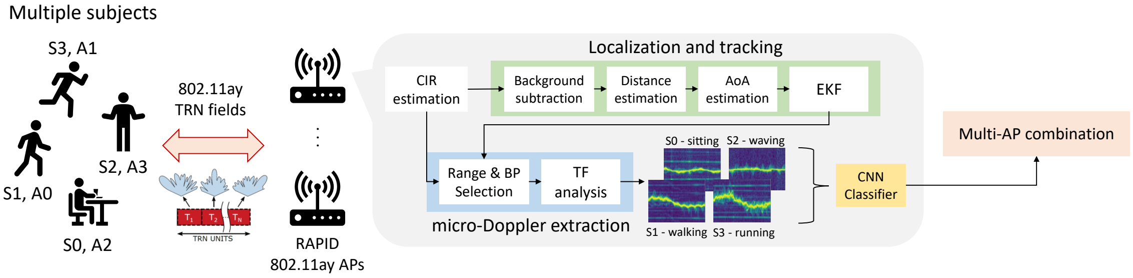


Figure 4.1: Overview of the RAPID system.

4.3.1 System overview

From a high-level perspective, the system performs the following operations, as shown in Fig. 4.1.
(1) IEEE 802.11ay CIR estimation: 802.11ay specifies the transmission of a variable number of TRN units for in-packet beam training, each using a (possibly) different Beam Pattern (BP). From the CIR, which is estimated from each TRN unit (see Section 4.4), RAPID obtains a scan of the whole angular FoV, which contains accurate information about all the surrounding objects and people.
(2) People localization and tracking: the individuals are detected by performing background subtraction from the CIR amplitude and applying a thresholding algorithm to detect candidate reflection paths from humans, see Section 4.3.3 and Section 4.3.3, respectively. Subsequently, a correlation based algorithm is utilized to estimate the angular position of the subjects, as described in Section 4.3.3, and an EKF is exploited to sequentially track and refine the positions of the individuals across time (Section 4.3.3). By combining more than one AP, RAPID can boost its human detection capabilities, while effectively coping with occlusion problems, as quantified in Section 4.6.2.
(3) μD spectrum extraction: here, the μD spectrum of each detected person is extracted. This is implemented by utilizing the CIR model as a radar return signal, and using the estimated positions from point (2) to single out the CIR portions (the paths and the BPs) containing the contributions of each subject, see Section 4.3.4. The μD signature of each individual's movement is then extracted by computing the power spectrum of the corresponding complex-valued portion of the CIR over windows of suitable length, employing Time-Frequency (TF) analysis.
(4) HAR and person identification: the spectrograms from step (3) are fed to a deep learning classifier based on a residual CNN [109] for HAR. Thanks to the separation of the CIR, and to the subsequent computation of the μD for each individual, RAPID is capable of recognizing the different activities performed by multiple subjects within the same indoor space. Moreover, through a second CNN module, it is also able to identify a person, by extracting and analyzing their gait features from the μD signature. With multiple APs, the classifications are refined by selecting the best AP to make the decision, according to the confidence of the classifier output.

In this chapter, we aim at localizing and tracking people within a given physical space, by identifying which person is performing which activity. This requires person identification, tracking and HAR capabilities. The person identification task is carried out by extracting and analyzing the μD associated with the human gait, as this is an effective (soft) biometric signature, which has been successfully used in many works [15]. Hence, we first detect when a person is walking, then we get his/her identity from the μD gait signature and, finally, we keep tracking the person by also recognizing their activities. This also works the other way around, i.e., if a person is at first sitting and doing other activities, and then starts walking later on; as long as tracking works, we can later determine who was sitting earlier on. This also explains why tracking a person is critical, so that it is still clear which person is where, even when he/she performs other activities than walking.

We now present in detail each RAPID processing function, following the workflow of Fig. 4.1.

4.3.2 CIR model

CIR estimation is a key component of most communication systems and is used to obtain information about the environmental reflections of the signal, such as their associated angle of arrival and delay at the receiver. A key aspect to our design is that the large transmission bandwidth of mmWave systems leads to CIR containing fine-grained information about the environment. In our system, the transmitter and the receiver units are co-located: the signal sent by the former, after bouncing off nearby reflectors (objects or humans), is collected at the receiver that retrieves information for each reflector, such as its distance and angle from the receiver, its velocity and micro-Doppler.

We consider a multipath propagation environment with a time-varying number of reflectors, $P(t)$. These cause physical signal propagation paths that can be separated in the CIR according to a finite *ranging resolution*, i.e., the capability of the system to resolve the distance of the reflectors causing different signal paths. This is given by $\Delta d = c/2B$, where B is the transmitted signal bandwidth and c the speed of light. Thus, the CIR contains the complex channel gains for a discrete grid of possible signal paths (or *distance bins*), with indices $\ell = 0, \dots, L-1$. These are obtained by correlating the received signal with pre-defined Golay sequences. Each path is associated with a specific distance from the AP, according to the relation $d_\ell = c\tau_\ell/2$, with τ_ℓ being the delay associated with path ℓ. The vector containing all the distances of interest is defined as $\mathbf{d} = [d_0, d_1, \dots, d_{L-1}]^T$. If multiple CIR estimations are performed over a single packet, using different BPs, the reflections from the environments are amplified differently. This is due to the different BP shapes, as each BP steers the transmission signal towards a specific direction (beam steering). Therefore, the CIR depends on the specific BP used during the transmission, denoted by $b = 0, \dots, N_{\text{BP}} - 1$. For carrier frequency f_o, the CIR along ℓ, using BP b at time t is

$$h_{\ell,b}(t) = \sum_{p=1}^{P_\ell(t)} a_{\ell,b}^p(t) \exp\left\{-j2\pi \frac{2f_o}{c}\left[d_\ell + \int_0^t v_\ell^p(x)dx\right]\right\}. \tag{4.1}$$

In Eq. (4.1), $P_\ell(t)$ is the number of physical reflectors whose contributions overlap in the ℓ-th

CIR path, as their distances are within $d_\ell \pm \Delta d/2$, while v_ℓ^p is the radial velocity[*] of reflector p. The quantity $a_{\ell,b}^p(t)$ is the complex gain due to the joint effect of the transmitter BP, the object reflectivity and the signal attenuation. For the sake of the content presented in this chapter, we can for the moment neglect the multiple propagation paths that overlap in a single distance bin, and assume that the summation in Eq. (4.1) contains only one term. This is a good approximation in mmWave systems, due to their high ranging resolution given by the high transmission bandwidth. Nevertheless, we will relax this assumption in Chapter 5, where we will consider and arbitrary $P_\ell(t)$ and show its impact on the μD extraction process. For this reason, in the following we refer to path ℓ as the single propagation path that can be detected in the ℓ-th CIR bin.

Eq. (4.1) is a continuous-time expression. In practice, the CIR estimation is repeated at discrete time instants that coincide with the reception of each packet, indicized by variable k. This can be seen as sampling the CIR in time, with sampling period corresponding to the inter-packet transmission time T_c. The expression of the ℓ-th path of the CIR obtained using beam-pattern b at time kT_c (packet k) is

$$h_{\ell,b}(kT_c) = a_{\ell,b}(kT_c) \exp\left\{-j2\pi \frac{2f_o}{c} \left[d_\ell + v_\ell^p(kT_c)kT_c\right]\right\} = a_{\ell,b}(kT_c)e^{j\phi_\ell(kT_c)}, \tag{4.2}$$

where $a_{\ell,b}(kT_c)$ and $\phi_\ell(kT_c)$ are the complex gain of path ℓ at time kT_c and its phase, respectively. The path gain depends on the contribution of the BP used for the transmission and on the reflectivity of the target, whereas the phase depends on the delay $\tau_\ell = d_\ell + v_\ell^p(kT_c)kT_c$. The movement speed of the reflector is considered to be time-varying, in Eq. (4.2), for the sake of presenting the most general CIR model.

4.3.3 People localization and tracking

CIR estimation is followed by people localization and tracking. This can be further split into *(i)* background subtraction, to remove the reflected paths due to static objects, *(ii)* the estimation of the subjects' distances, *(iii)* the estimation of the angular positions of the subjects with respect to the device, and *(iv)* their joint processing using a Kalman filter to track each person's trajectory across time.

HAR and identification require CIR readings at a rate $1/T_c$, whereas localization and tracking use a time granularity of $\Delta t > T_c$ seconds, where index r denotes the localization/tracking time-steps. The choice of setting $\Delta t > T_c$ stems from the fact that performing localization and tracking for every transmitted packet is unnecessary, as the packet transmission rate $1/T_c$ is much larger than the speed of human motion. So, the system computes estimates at different rates, according to the specific resolution that is required by each. This allows for an additional flexibility in the selection of the type of BPs that are used for each packet: as we explain shortly below in Section 4.4 and Section 4.5, we can modulate how many TRN units are included in a packet according to the type of sensing function that is being performed, i.e., localization/tracking versus activity/identity recognition.

[*] By convention, v_ℓ^p has a positive sign when the reflector moves away from the AP.

Background subtraction

To infer the positions of the subjects it is key to remove the reflections due to static (background) objects, as these typically have a much higher intensity than those generated by humans and may impact the localization accuracy. The background-related CIR is estimated by computing the time average of the CIR amplitude within a window of K_{static} samples, as static reflections are constant across time,

$$\bar{h}_{\ell,b} = \frac{1}{K_{\text{static}}} \sum_{k=0}^{K_{\text{static}}-1} |h_{\ell,b}(kT_c)|. \tag{4.3}$$

Then, the foreground CIR amplitude component is obtained as $|\tilde{h}_{\ell,b}(r)|= \max\left(|h_{\ell,b}(r)|-\bar{h}_{\ell,b}, 0\right)$, i.e., removing the amplitude of the static paths and setting to zero the amplitude of those paths that would be present in the reference background CIR, but that are shielded by the presence of a person. We remark that, through different BP, we perform beam steering at the transmitter. Hence, the peaks in $|\tilde{h}_{\ell,b}|$ correspond to the strongest propagation paths, as seen at the receiver when beam-pattern p is used at the TX side. Changing the BP p allows scanning the environment by varying the transmission angle and, in turn, sweeping the whole field of view. We use this to infer the distance and the angular position of each individual, as described next.

Distance estimation

The distance of each subject is obtained by applying a threshold on $|\tilde{h}_{\ell,b}|$ (the time index is omitted for better readability), selecting the strongest paths across all the used BPs. First, for each reflected path ℓ, we consider vector

$$\mathbf{h}_\ell = \left[|\tilde{h}_{\ell,0}|, |\tilde{h}_{\ell,1}|, \dots, |\tilde{h}_{\ell,N_{\text{BP}}-1}|\right]^T, \tag{4.4}$$

containing the CIR values of path ℓ for each of the N_{BP} BPs that are used at the transmitter. We collect the L_2-norms of $\mathbf{h}_\ell$, with $\ell = 0, 1, \dots, L-1$, obtaining a new vector $\mathbf{h}$, as

$$\mathbf{h} = \left[||\mathbf{h}_0||_2, ||\mathbf{h}_1||_2, \dots, ||\mathbf{h}_{L-1}||_2\right]^T, \tag{4.5}$$

containing the strengths of each path at the receiver. We locate the local maxima in $\mathbf{h}$, denoting them by $h'_0, h'_1, \dots, h'_{\text{n}_{\text{peaks}}-1}$. Hence, we discard those peaks with amplitude smaller than a dynamic threshold A_{th} computed from the maximum and average power of the paths in the current CIR. We introduce the following coefficients $\alpha_{\max}$, α_{mean} and α_{abs}, and compute the threshold value A_{th}, as

$$A_{\text{th}} = \max\left\{\alpha_{\max} \cdot \max_i h'_i, \alpha_{\text{mean}} \cdot \bar{h}', \alpha_{\text{abs}}\right\}, \tag{4.6}$$

with $\bar{h}' = \sum_i h'_i/n_{\text{peaks}}$. We empirically assessed that suitable values for the coefficients are $\alpha_{\max} = 0.25$, $\alpha_{\text{mean}} = 2$ and $\alpha_{\text{abs}} = 2.5 \cdot 10^{-3}$. With Eq. (4.6) the threshold is computed dynamically, proportionally to the maximum between the average and the maximum value of the

CIR, while α_{abs} represents the minimum value we allow A_{th} to assume. The peaks that exceed the threshold are selected as candidate targets of interest and used for the subsequent AoA estimation. Denoting by $\ell_1, \ell_2, \dots, \ell_{N_s}$ the indices of the selected (candidate) paths ($0 \leq \ell_j \leq L-1$), the corresponding distances are obtained as $d_{\ell_j} = c\tau_{\ell_j}/2$.

Angular position estimation

The following procedure is applied to each of the N_s candidate paths. Let vector $\mathbf{s}_{\ell_j} \in \mathbb{R}^{N_{\mathrm{BP}}}$ contain the squared CIR amplitudes from one of such paths, ℓ_j, for all used beam patterns, i.e., $\mathbf{s}_{\ell_j} = \left[|\tilde{h}_{\ell_j,0}|^2, |\tilde{h}_{\ell_j,1}|^2, \dots, |\tilde{h}_{\ell_j,N_p-1}|^2\right]^T$. $\mathbf{s}_{\ell_j}$ is normalized by dividing it by its L_2-norm $||\mathbf{s}_{\ell_j}||_2$, then a correlation measure is used to estimate the angular position of the target by exploiting the gains of each beam pattern along the azimuth angular FoV θ. Specifically, denoting by $g_b(\theta) \in [0,1]$ the normalized gain of beam pattern p along direction θ (see Fig. 4.5b), the angular position for candidate path ℓ_j is estimated as

$$\theta_{\ell_j} = \operatorname*{argmax}_{\theta} \sum_{b=0}^{N_{\mathrm{BP}}-1} g_b(\theta) \frac{|\tilde{h}_{\ell_j,b}|^2}{||\mathbf{s}_{\ell_j}||_2}. \tag{4.7}$$

The rationale behind Eq. (4.7) is that if $|\tilde{h}_{\ell_j,b}|$ originates from the signal reflected off a subject, the corresponding angular direction is the one leading to the highest correlation between the CIR squared amplitude and the set of beam pattern gains. This is because each BP amplifies path ℓ_j differently, depending on the beam pointing direction.

Upon obtaining the distance and the angle estimates, an Extended Kalman filter is utilized to track the subjects' positions over time.

People tracking - extended Kalman filter

After the localization step, the candidate positions of the subjects are known in polar coordinates, and constitute our *observations* of the positions of the subjects, which we denote by $\mathbf{z}_r^j = [d_{\ell_j}, \theta_{\ell_j}]^T, \forall j = 1, 2, \dots, N_s$. We employ an EKF [67] to track the physical position of each individual in the Cartesian space. Specifically, we define the true state of subject j at time r as vector $\mathbf{x}_r^j = \left[x_r^j, y_r^j, \dot{x}_r^j, \dot{y}_r^j\right]^T$, containing the coordinates along the $x-y$ horizontal plane and the movement velocity components along the same axes. We approximate the motion of the subjects with a constant velocity (CV) model [110]. As the observations $\mathbf{z}_r^j$ become available, we apply the predict and update steps of the EKF to follow the movement trajectories of the subjects [67]. The association between the observations from time $r+1$ and the states from time r is done using the nearest-neighbors joint probabilistic data association algorithm (NN-JPDA) [89].

Using the EKF estimates $\hat{\mathbf{x}}_r^j$ of each person's state across subsequent time steps allows retrieving the path and the BPs in the CIR which contain his/her μD signature.

4.3.4 micro-Doppler extraction

CIR phase model

The CIR model in Eq. (4.2) is here expanded and related to radar theory [11]. Using a typical radar terminology, we refer to the CIR samples $\ell = 0, 1, \dots, L-1$ as the *fast-time* sampling dimension, as they are obtained at the highest available sampling rate. The CIR samples collected across different packets are instead referred to as the *slow-time* samples, indicized by variable k as in Section 4.3.2.

Next, we consider a moving object within the monitored indoor space; the transmitted signal is reflected off the object and the corresponding contribution is retrieved at the receiver in the ℓ-th path of the CIR. To extract the μD effect caused by the movement of this object, we analyze the phase of the ℓ-th path across time. The time-dependent phase term in Eq. (4.2) can be expressed as follows

$$\phi_\ell(kT_c) = -2\pi f_o \frac{2\left(d_\ell - v_\ell(kT_c)kT_c\right)}{c} \approx -2\pi f_o \bar{\tau}_\ell + 4\pi f_o \frac{v_\ell}{c} kT_c. \tag{4.8}$$

Here, $\bar{\tau}_\ell$ is the delay of the ℓ-th path due to the distance of the corresponding reflector from the device. v_ℓ is the radial velocity of the reflector with respect to the device, which is here assumed to be *slowly* time-varying, i.e, we can consider approximate it as constant during a μD spectrum processing interval (see Section 4.3.4). From Eq. (4.8) it can be seen that the velocity of the object at distance d_ℓ, if greater than zero, modulates CIR phase across the slow time dimension. Following a common convention [11], in this chapter objects moving away from the transmitter (AP) have *positive* velocity, while incoming objects have *negative* velocity.

The human body contains multiple moving parts that have different velocities and follow different trajectories. Thanks to the small wavelength of mmWave, in the μD we can observe these different contributions via TF analysis, as detailed in the next Section 4.3.4.

micro-Doppler spectrum

Human movement causes a frequency modulation on the reflected signal due to the small-scale Doppler effect produced by the different body parts. Using TF analysis of the received signal, it is possible to distinguish between different actions performed by a person or identify the individual based on his/her way of walking (*gait*) [7], [23]. mmWave radios are particularly suited for this, as their frequencies are sensitive to the μD effect due to their small wavelengths.

From Eq. (4.8), the μD effect of human movement can be extracted from subsequent estimates of the CIR, computed every T_c seconds. Specifically, one can compute the STFT of $h_{\ell,b}(kT_c)$, across slow-time, for each path ℓ and each beam pattern p as

$$H_{\ell,b}(m,i) = \sum_{l=0}^{W-1} h_{\ell,p}(l + m\delta) w(m) e^{-j2\pi \frac{il}{W}}, \tag{4.9}$$

where m is the time index, $i = 0, 1, \dots, N_D - 1$ is the frequency index, W is the (fixed) window length, w is a Hann window of dimension W and δ is the time granularity of the STFT. In Eq. (4.9) we omitted the sampling time T_c for readability, as it is constant ove the whole window. The power spectrum of $h_{\ell,b}(kT_c)$, computed as $\mu_{\ell,b}(m,i) = |H_{\ell,b}(m,i)|^2$, contains information on the phase modulation due to the velocity v_ℓ, and can be used to analyze its evolution across subsequent windows.

Eq. (4.9) can not be used directly to extract the μD signature of a moving human in our setup, as it refers to a single fast time bin (a single path in the CIR) and a single BP, while people can be located in different positions across time. In addition, it would be inefficient to compute the STFT for all the paths and all the BPs. Instead, the computation should only be performed for those physical locations where a person is detected. In the following, we leverage the localization and tracking process described in Section 4.3.3 to only extract the CIR portions that contain useful μD information.

μD separation

Assume that we want to extract the μD of a person that was detected and located by the previous algorithms at a certain distance and angle with respect to the device. Hence, we extract the CIR samples from the most useful BP, i.e., the one that points in the direction of the person and that, in turn, emphasizes the most the reflection from this target.

From the estimated state of this person (Section 4.3.3), his/her angular position is obtained as $\hat{\theta} = \arctan(\hat{y}_r/\hat{x}_r)$ and his/her distance from the device, as $\hat{R} = \sqrt{\hat{x}_r^2 + \hat{y}_r^2}$. The BP approximately pointing in the direction of this person, denoted by b^*, is thus selected as the BP having the highest gain along $\hat{\theta}$, that is

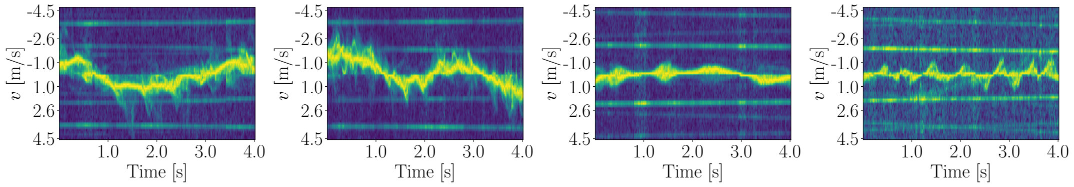
$$b^* = \underset{p}{\operatorname{argmax}}\, g_b(\hat{\theta}). \tag{4.10}$$

Moreover, due to the high ranging accuracy of mmWave, humans typically produce reflections that influence more than a single CIR path. The CIR paths of interest are those that correspond to a neighbourhood of $\hat{R}$. In our analysis, we take the size of this neighborhood constant across all subjects, denoting it by Q. Specifically, we first select the path ℓ^* that best matches the subject's distance $\hat{R}$

$$\ell^* = \underset{\ell}{\operatorname{argmin}} |d_\ell - \hat{R}|. \tag{4.11}$$

Then, from the original complex-valued CIR, we extract a window containing Q samples along the fast-time dimension, centered on ℓ^*. For convenience, we assume Q to be an odd integer, as this makes the following processing steps symmetric with respect to a central CIR path (corresponding to the torso), but the same steps can be applied for Q being even. We aggregate the spectra obtained from the path caused by the torso, ℓ^*, with the $\lfloor Q/2 \rfloor$ distance bins preceding ℓ^* and the $\lfloor Q/2 \rfloor$ subsequent distance bins, as they may contain the contributions of the other body parts. The expression of the i-th μD spectrum component is

$$D_i(m) = \sum_{\ell=\ell^*-\lfloor Q/2 \rfloor}^{\ell^*+\lfloor Q/2 \rfloor} |H_{\ell,b^*}(m,i)|^2, \quad i = 0, 1, \ldots, N_D - 1, \tag{4.12}$$



Walking. Running. Sitting down. Waving hands.

Figure 4.2: Example 4 s long μD spectrograms obtained by RAPID from 4 subjects. The yellow and blue colors respectively represent high and low power in the corresponding Doppler velocity bins (y axis).

while the total spectrum is represented by vector $\mathbf{D}(m) = [D_1(m), D_1(m), \ldots, D_{N_D-1}(m)]^T$ To capture the human movement evolution across time, we compute the μD vectors for a window of $N_{\mu\text{D}}$ subsequent time-steps and concatenate them into a spectrogram representing the μD signature of the target up to time m, as

$$\boldsymbol{\Upsilon}_m = [\mathbf{D}(m - N_{\mu\text{D}} + 1), \mathbf{D}(m - N_{\mu\text{D}} + 2), \ldots, \mathbf{D}(m)] . \tag{4.13}$$

The procedure described in this section is repeated for all the detected subjects.

Human μD range and resolution

During everyday movement, the limbs of a person usually have velocities of up to $3 - 4$ m/s [7], [23]. To fully capture the μD signature of the subjects, we must ensure that our systems achieves a sufficient resolution. Recalling Eq. (4.8), we know that the Doppler frequency shift induced by a moving object on the ℓ-th path is $f_\ell^\text{D} = 2f_o v_\ell / c$. Using TF analysis to estimate the Doppler spectrum as in Eq. (4.9), the resolution that can be obtained on the Doppler frequency is $\Delta f^\text{D} = 1/(WT_c)$. The maximum measurable Doppler frequency is instead $f^\text{D}_\text{max} = 1/(2T_c)$. These quantities can be mapped onto the velocity estimate resolution and the maximum measurable velocity as

$$\Delta v = \frac{c}{2f_o W T_c}, \quad v_\text{max} = \frac{c}{4 f_o T_c}. \tag{4.14}$$

Given that we sample the CIR on a per-packet basis, to capture the μD effect of human motion we must ensure that the time T_c between the packets used in the μD estimation allows capturing the range of velocities of interest. See also Section 4.6 for the chosen values of W and T_c.

4.3.5 Activity recognition and person identification

The μD signature, obtained as in Eq. (4.13), contains information about the type of movement performed by the person.

To perform HAR and person identification, we use a deep neural network to classify each spectrogram. Specifically, once the μD signatures of each person have been separated, RAPID performs the following tasks: *(i)* it classifies the activity carried out by the subject into *walking* (A0), *running* (A1), *sitting down* (A2), *waving hands* (A3) and *standing still* (A4) and *(ii)* it recognizes the subject's identity during a walking phase, among a known set of individuals, denoted by S0, S1, etc. In Fig. 4.2, we show μD signature examples for activities A0 $-$ 3, concurrently performed by 4 subjects within the same environment.

As human μD is highly variable across different subjects, and we seek robustness to different environment conditions and noise, we employ deep learning to classify the μD signatures. Referring to a single subject, the μD spectrum $\mathbf{\Upsilon}_m$ is represented as an image and processed by two separate CNN for HAR and person identification, respectively. The two classifiers share the same architecture, as shown in Fig. 4.3, but are trained separately and have different weights as they perform different tasks. As the subjects are continuously tracked over time, we adopt a sliding window approach, selecting μD spectrograms with $N_{\mu\mathrm{D}}$ μD spectrum samples for each window (matrix $\mathbf{\Upsilon}_m$). Subsequent windows partially overlap to increase the reactivity of RAPID in obtaining predictions. Both CNN are trained to extract features from the μD spectrograms and to classify the activity performed by or identity of the person, by learning a function $\mathcal{F}(\cdot)$ that maps a μD window, $\mathbf{\Upsilon}_m$, of size $N_D \times N_{\mu\mathrm{D}}$, onto a vector $\mathbf{c}_n$ containing the HAR (identification) class probabilities, i.e., $\mathbf{c}_m = \mathcal{F}(\mathbf{\Upsilon}_m)$. The dimension of the final probability vector $\mathbf{c}_m$ is different in case of HAR or identification depending on the dimension of the classification problem. The second CNN, used for person identification, is only trained on walking spectrograms, as human gait is well known to be a soft biometric identifier [15]. Hence, during the system operation, the identification classifier is only applied on the input μD spectrogram when the activity is classified as "walking" by the HAR classifier, see Fig. 4.3.

μD spectrogram pre-processing

Prior to feeding it to the CNN classifier, the μD spectrogram is pre-processed by removing the contributions from static reflections and normalizing it.

Static reflection removal. A customary step when processing human μD signatures is the removal of static reflections, which appear as a strong power peak around the 0 m/s velocity bin. This can be done by either applying a high-pass filter to the signal or, if deep learning methods are used for classification, by directly removing the Doppler bins containing unwanted contributions, as done in [5], [7]. We adopt the latter method to remove the Doppler bins corresponding to the velocities in the interval $[-0.28, 0.28]$ m/s, as they contain very low, non-informative velocities.

Normalization. To compensate for differences in the strength of the reflections when subjects are far from the APs, we normalize each column of $\mathbf{\Upsilon}_m$, $\mathbf{D}(j), j = 0, 1, \dots, N_{\mu\mathrm{D}} - 1$ in the range $[0, 1]$,

$$\mathbf{D}(j) \leftarrow \frac{\mathbf{D}(j) - \min_i \mathrm{D}_i(j)}{\max_i \mathrm{D}_i(j) - \min_i \mathrm{D}_i(j)}. \tag{4.15}$$

Deep learning classifier

We use the same CNN architecture, based on deep residual networks [109], for HAR and person identification, with the only difference being the dimension of the last classification layer. This network consists of 4 consecutive residual blocks. Each residual block has two convolutional layers

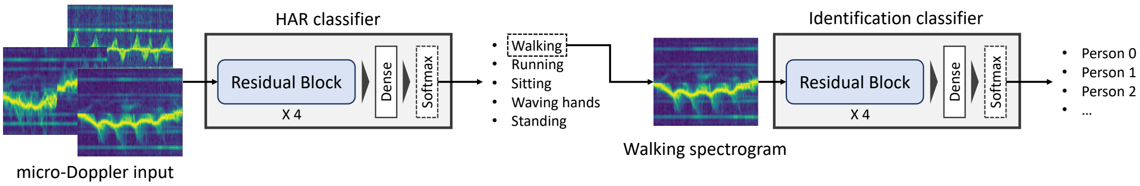


Figure 4.3: Block diagram of the CNN classifiers used by RAPID for HAR and person identification.

[53], the first of which includes a down-sampling by a factor of 2 (*stride*). Each convolution is followed by an ELU activation function [54] and batch normalization [55]. The output of the convolution is summed to the input (*skip connection*) and passed through another ELU activation and batch normalization. The 4 residual blocks use 8, 16, 32 and 64 filters, respectively, all having a kernel of size 3×3. After the last residual block, we apply Dropout [57] with a ratio of 0.5, and a fully-connected (or *dense*) layer with 64 units, then, a second Dropout operation with ratio of 0.2. Finally, the classification probabilities for HAR or person identification are computed via a Softmax activation function [53]. The network architecture is shown in Fig. 4.3.

Combining multiple APs

Using the different points of view provided by the different APs, RAPID can improve its HAR and person identification performance. Assume that a person is independently detected and tracked by 2 or more APs concurrently. A slightly different μD signature of the person is obtained by each AP, according to the angular position and the distance of the device with respect to the person. At each time instant m, we adopt a simple decision fusion scheme including the following steps: *(i)* if a single AP detects the person, the decision made by the classifier on the corresponding μD signature is used, i.e., $\text{argmax}_j\, c_{m,j}$, where $c_{m,j}$ is element j of vector $\mathbf{c}_m$, *(ii)* if multiple APs detect the person, denote by $\mathbf{c}_m^a$ the probability vector predicted by AP a. The final decision is made by the AP that is most confident about its classification, i.e., the one that assigns the highest probability to the predicted class: $\text{argmax}_j \left\{\max_a c_{m,j}^a\right\}$.

4.4 Enabling sensing in IEEE 802.11ay

The high bandwidth of IEEE 802.11ay [111] not only provides high data throughput but also offers excellent accuracy for sensing applications. RAPID is able to extract highly accurate range, angle and μD information from CIR measurements. For this, we take advantage of the beam training and beam tracking mechanisms of IEEE 802.11ay systems.

Range and angle information are extracted from the CIR obtained via the Channel Estimation Field (CEF) of standard beacon frames that are frequently sent by the AP or the beam training frames sent during a Sector Level Sweep (SLS). The SLS is a two-step procedure: first, one device sends training frames using the available antenna configurations, while the second device listens using a quasi omnidirectional BP. Then, the devices exchange their roles to train the other device. After sending feedback, the devices can select the *best* combination of BP on both sides of the

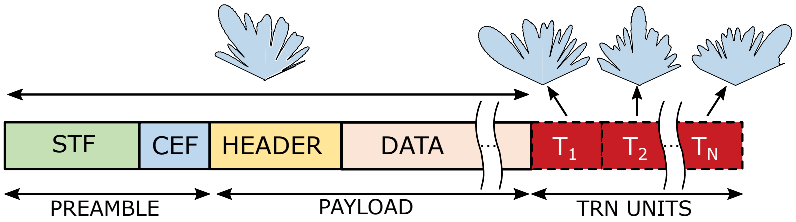


Figure 4.4: IEEE 802.11ay *in-packet* TRN fields.

link. IEEE 802.11ay also introduces the concept of *in-packet* beam tracking [25], where different antenna configurations can be tested within a single packet, allowing for much quicker BP changes. This is done by appending a TRN field to the packet as shown in Fig. 4.4. A TRN field is composed of multiple (variable) TRN units formed by 6 complementary Golay sequences of type a ("Ga") and b ("Gb") with length 128 samples each:

$$\{+\text{Ga}_{128};\ \text{-Gb}_{128};\ +\text{Ga}_{128};\ +\text{Gb}_{128};\ +\text{Ga}_{128};\ \text{-Gb}_{128}\}\,. \tag{4.16}$$

The excellent autocorrelation properties of the complementary Golay sequences and the availability fast hardware structures for the correlation [112] make them ideal for CIR estimation [26]. Indeed, the sum of the autocorrelation sequences of a pair of complementary Golay sequences, $\{\text{Ga}_{128};\ \text{Gb}_{128}\}$ gives exactly a Kronecker delta function, without sidelobes [113]. This makes them suitable for channel estimation in multipath environments such as indoor scenarios. The high bandwidth (1.76 GHz) of the IEEE 802.11ay channels results in a range resolution of $\sim$ 8.5 cm directly from the CIR estimate. Denoting by ℓ a delay bin, by $h^i_{x,\ell}$ the CIR for the i-th pair of complementary Golay sequences with $x \in \{\text{a}, \text{b}\}$, by R the TRN field of the received packet, we have

$$h^i_{x,\ell} = \sum_{n=0}^{127} R(\ell+n)\text{G}x^*_{128}(n). \tag{4.17}$$

Then, the final CIR estimate is obtained summing the complementary pairs and averaging over all the repetitions contained in a TRN field as

$$h_\ell = \sum_{i=1}^{3} h^i_{\text{a},\ell} + h^i_{\text{b},\ell}. \tag{4.18}$$

The above equations are used to estimate the CIR for every TRN field, thus we omitted the indices referring to the packet index (time), k, and to the BP, b. When referring to a specific k and b, Eq. (4.18) represents the CIR modeled in Eq. (4.2). Considering the different BP shapes used during beam training, possible targets located in the FoV of the devices are *illuminated* by the respective BPs that focus energy in that direction and they appear as multi-path components in the CIR (Fig. 4.5d). Furthermore, we take advantage of the different amplification factors in the multi-path components (given by the different BPs) to estimate the angular positions of the subjects. For this purpose, we apply the correlation based approach explained in Section 4.3.3 to the different channel multi-path components in the channel. Considering the common speed of human motion, carrying out beaconing or a beam training procedure every, e.g., 100 ms allows accurately locating human targets in the FoV of the AP. Note that, as we show in Section 4.6, a *full* beam training, that scans all the available BP, is in fact not needed, and we may use a much smaller subset of BP.

Extracting μD signatures from the CIR requires fine-grained frequency resolution, as detailed in Section 4.3.4. This cannot be achieved with the CIR estimates obtained from beacons or beam training packets only, as sampling the CIR with $T_c = 100$ ms would lead to an insufficient maximum measurable Doppler velocity of $6.25 \cdot 10^{-3}$ m/s (see Eq. (4.14)). We address this by exploiting the *beam tracking* procedure defined in the standard [111]. It allows to add a configurable number of TRN units to data packets to test different BP configurations to *quickly* correct possible misalignment without requiring a full beam training procedure.

After identifying the subjects' ranges and angles using beam training packets, we include a TRN field in *data packets* with a sufficient number of TRN units to illuminate all the subjects in the scene; each TRN unit uses a suitable BP that specifically points in the direction of a person. This steers the energy of the transmitted signal so as to best capture the μD signatures of the subjects. Considering that data packets are sent much more frequently than beam training packets, our approach can sample the CIR with a sufficiently low T_c to capture the desired range of frequencies for human movement analysis.

4.5 Implementation

The available mmWave Commercial-Off-The-Shelf (COTS) devices support IEEE 802.11ad and offer very limited access to physical layer information [114]. To the best of the authors' knowledge, there are no COTS solutions for the new IEEE 802.11ay standard available yet. To address the lack of hardware, we turn a mmWave SDR system into a ISAC experimentation platform. Here we cover the design decisions made to implement RAPID on such platform.

4.5.1 Hardware components

As a baseline to implement a RAPID AP, we use the mm-FLEX experimental platform [115]. This open platform is composed of a baseband processor including a Xilinx Kintex Ultrascale FPGA plus high-speed AD/DA converters and DDR memory banks. Besides, it is connected through a PCIe interface to a Core i7 processor card co-located within the same hosting chassis. The latter implements configuration and control tasks not only for the FPGA and converters, but also for the RF front-end.

The baseband processor is configured to fulfill the bandwidth requirements of IEEE 802.11ad/ay standards (1.76 GHz), using a sampling frequency of 3.52 GSPS for both AD/DA converters, with 2 samples per symbol.

The RF front-end includes a 60 GHz up/down converter and phased antenna arrays from Sivers [116]. The device is able to operate on all the channels defined in the IEEE 802.11ad/ay standards

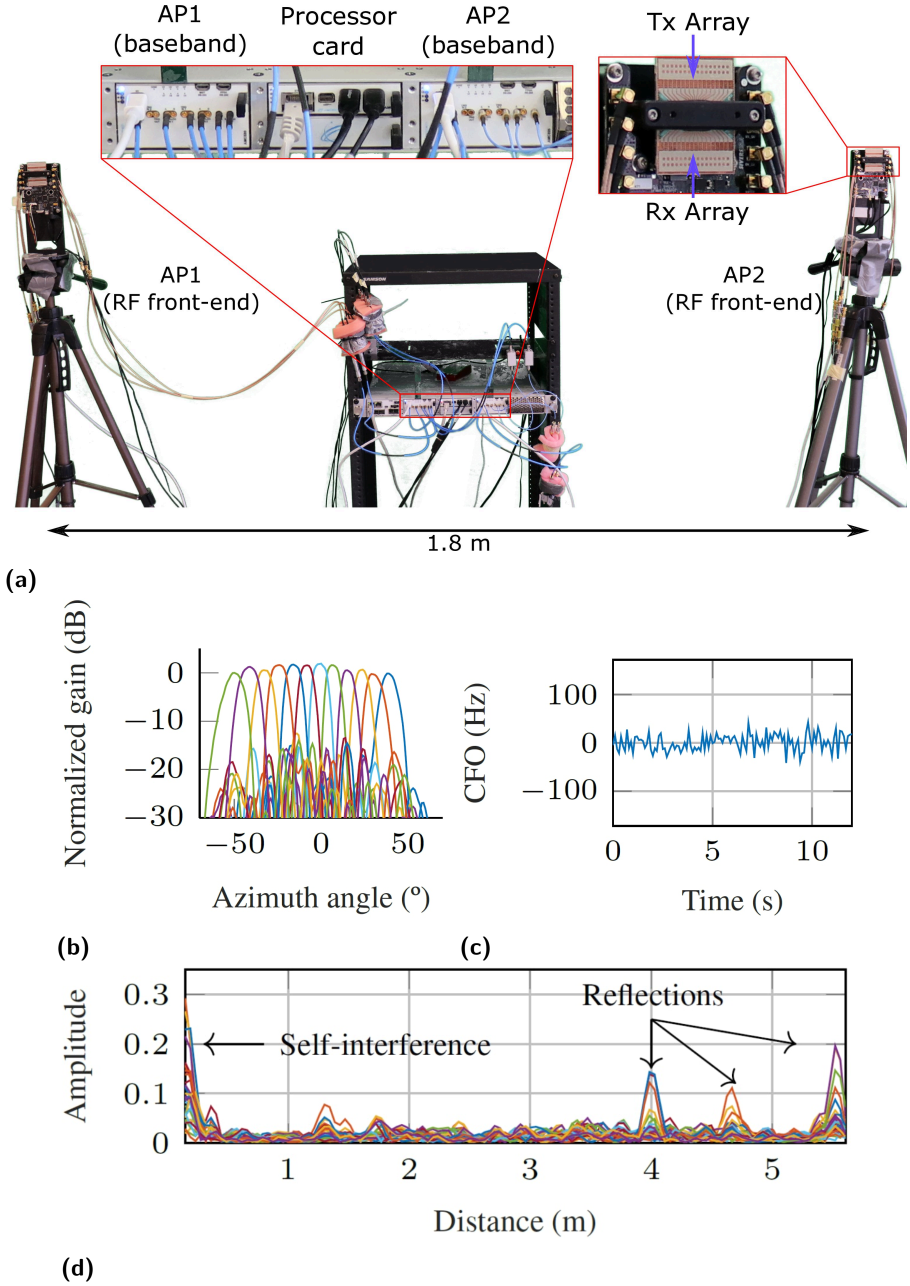


Figure 4.5: RAPID implementation.

[111], [117]. As shown in Fig. 4.5a, the device integrates two independent 16-element linear antenna arrays, one used for transmission and one for reception. The codebook of BP for both arrays can be freely configured. The system is controlled in real-time using USB and SPI interfaces, as well as GPIO pulses for the quick BP changes required for beam training and tracking.

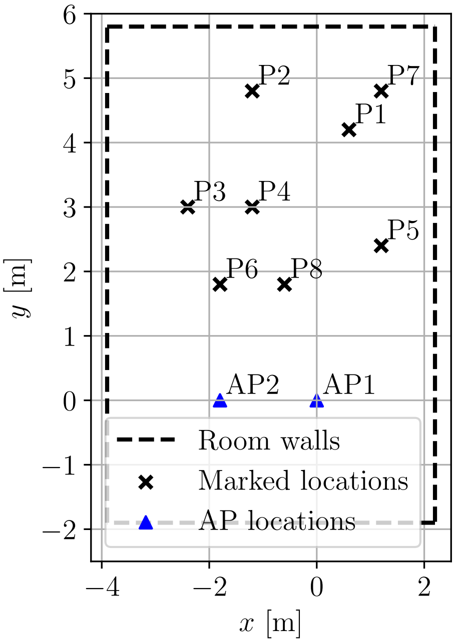


Figure 4.6: Schematic representation of E1.

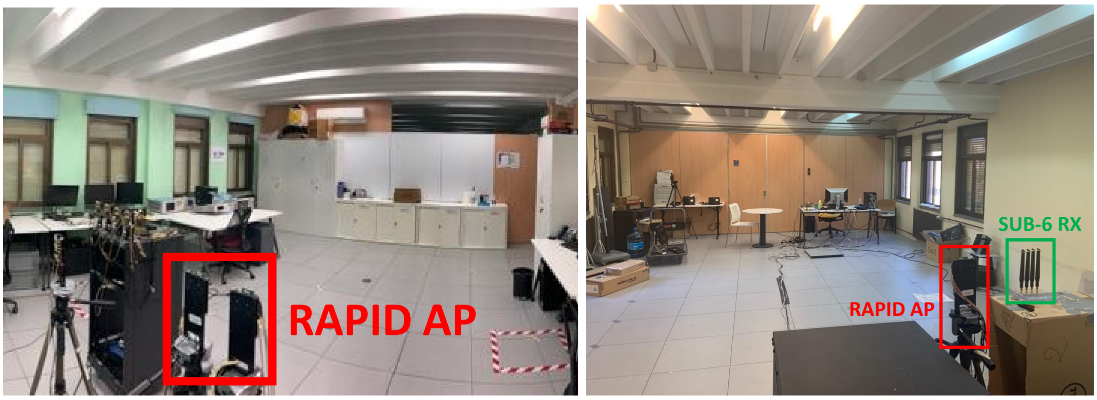


Figure 4.7: The two environments: E1 (left) and E2 (right).

4.5.2 Full-duplex operation

To bring radar capabilities to the experimentation platform, it is necessary to support simultaneous operation of the TX and RX chains. This is achieved by concurrently enabling transmit and receive sub-systems in the RF front-end, and by enhancing the functionality of the baseband processor.

The 60 GHz front-ends used in this chapter [116] are laboratory equipment designed for early stage proof-of-concept communication systems. The carrier frequency is generated from a 45 MHz clock, which introduces significant Carrier Frequency Offset (CFO) and destroys the phase coherence between the CIR estimates obtained from consecutive packets. This would make the extraction of μD signatures infeasible with two independent co-located antennas. Instead, by using both transmit and receive arrays from the same RF front-end (see Fig. 4.5a), up and down conversion sub-systems are fed by the same local oscillator which keeps CFO levels in the range of $[-40, 40]$ Hz, as shown in Fig. 4.5c. Although transmit and receive arrays are directly next to each other, *no complex analog or digital self-interference cancellation techniques are required.* Thanks to the directional BPs and the robustness of the Golay Sequences of the TRN units, the system only requires some transmit power control to avoid saturating the receiver antennas and down-conversion stages. In Fig. 4.5d, we show the CIR measurements obtained from multiple BPs within a packet, by marking the self interference path and the reflections from the test room, where the different amplitudes correspond to the different BP shapes towards the direction of the reflectors.

In the baseband processor, we implement a state-machine on the FPGA logic which controls the transmitter and receiver data-paths. Specifically, it handles the DDR memory that stores the transmitted frames, performs multiple real-time antenna re-configurations over the TRN field of the packet, triggers the DDR memory on the receiver data-path, and sets the inter-frame spacing between multiple transmitted packets. While here we focus on an AP-centric design, the same procedure can be applied to implement RAPID on any station in the network.

Since our RAPID AP operates in a mono-static configuration, we perform CIR extraction without requiring the use of packet detection and synchronization circuits. To do this, it is important to ensure deterministic latency between the transmitter and receiver data-paths. Considering that transmit and receive data-paths have their own independent clock structure, we use clock-domain crossing techniques to send the state machine signals across transmit and receive domains. Besides, latencies in the DDR controllers are variable, which requires the use of FIFO queues at the output/input of the TX/RX DDRs. Together, these solutions help to achieve the desired deterministic latency.

4.5.3 Multi-AP system

Since IEEE 802.11ay networks typically involve many AP and dense deployments, we extend the aforementioned testbed capabilities to handle multi-AP scenarios. To this end, we integrate a second baseband processor in the hosting chassis which is connected to an independent 60 GHz front-end. The FPGAs from both processors have their own clocking structure, i.e., they are not synchronized. Each AP can be freely configured with its own parameters. For the sake of simplifying the system management, we use different communication channels (58.32 and 60.48 GHz) for each RF front-end, avoiding cross interference. It is worth mentioning that the channels can be freely configured, making it possible to operate the two RAPID APs so that they share the same frequency band, by implementing carrier sensing mechanisms.

4.6 Experimental results

In this section, we discuss the results of our extensive measurement campaign. Motivated by the discussion in Section 4.3.4 and Section 4.4, for the μD estimation we consider data packets (with TRN fields) spaced by $T_c = 0.27$ ms. This allows capturing velocities in the range $[-4.62, 4.62]$ m/s and leads to a resolution of $\Delta v = 0.14$ m/s when using a window of $W = 64$ samples in the DFT computation, see Eq. (4.14). These values are comparable to the ones achieved with radar devices [5], [7], [8]. Note that the even spacing of packets is just for convenience but is not a requirement, i.e., estimation can be done with random bursts of data packets with sufficiently small spacing. Moreover, we set to $Q = 9$ the size of the fast time window used to capture the contribution of the subjects in the CIR (see Section 4.3.4). The EKF time-step duration is set to $\Delta t = 32T_c$, which is also the time-granularity at which we obtain μD spectrum vectors. To extract range and angle information, we use in-packet beam training frames with 12 TRN units, using antenna beams covering a FoV range from $-45°$ to $45°$. With this configuration we achieve a mean accuracy of $2°$ for the angular position of a person standing in the room. We verify that this allows tracking multiple subjects reliably and with low localization error, as detailed in the following. In order to implement the angle estimation method from Section 4.3.3, we measured the BP shapes from the codebook using a motorized pan-tilt platform. In Fig. 4.5b, we show the 12 BPs we used to perform the experiments.

4.6.1 Experiment setup

We test RAPID in two different rooms, as shown in Fig. 4.7. The two environments are research laboratories, denoted by E1, of dimensions 6.1 × 7.7 m and E2, of dimensions 6 × 10.7 m (E2), and containing whiteboards, windows, tables, computers and equipment, making them challenging multi-path environments with a number of potential reflectors. Most of our experiments, including the collection of the training data for the NN classifier, have been carried out in E1, while we used E2 to assess the robustness of the proposed method to unknown environments. For the tests involving multiple AP, we deploy two RAPID APs as shown in Fig. 4.5a close to the wall, separated by 1.8 m.

To test the localization and tracking capabilities of RAPID, we mark specific known positions across E1 to determine the ground truth location as shown in Fig. 4.6, and perform our tests by having subjects move across these positions. The markers are denoted by Px, with x ranging from 1 to 8, while APs are represented as blue triangles. The room walls are represented with a black dashed line.

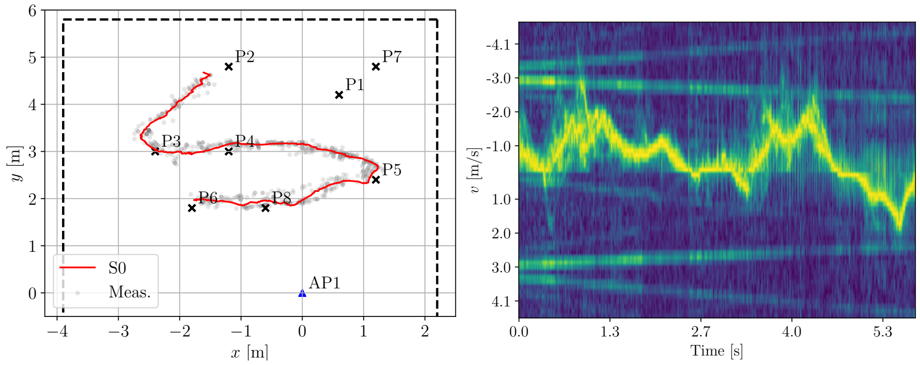


Figure 4.8: Subject walking trajectory (left) and a portion of the corresponding μD signature (right) extracted by RAPID.

4.6.2 Baseline experiments

We first report the results obtained in two simple baseline experiments to verify the capability of RAPID to extract the μD signature of a moving person in an indoor scene. Here, we only use AP 1 and a single subject, performing different activities at different locations in E1.

Fig. 4.8 shows the EKF estimated trajectory of the subject walking along the trajectory P2-P3-P4-P5-P8-P6 together with the corresponding μD spectrogram. The light grey points represent the raw measurements (observations) obtained as explained in Section 4.3.3, using Cartesian coordinates. The trajectory is correctly reconstructed with remarkable accuracy. The μD signature is extracted successfully and shows the different contributions of the torso and the limbs. The former reflects more power and follows a slightly oscillating motion, which is coherent with the direction changes in the walking trajectory, while the latter are responsible for the higher velocity peaks.

Next, we test RAPID on a subject sitting down at the marker P2, as shown in Fig. 4.9. Also in this case, RAPID correctly estimates the location of the subject, and the μD spectrum is coherent with the sitting down activity. This is non-trivial, given that P2 is located at the edge of the experiment room. The empirical CDF of the positioning error of the subject in Fig. 4.10 shows that we achieve a good localization accuracy. In this analysis, we included around 2000 position estimates made by the EKF. The median error is 26 cm, and the probability of the error being lower than 40 cm is close to 1. We stress that the subject in this case is not static, as the person alternates between sitting down and standing up. This causes the estimated position to change slightly across time-steps, increasing the localization error.

4.6.3 Multi-person multi-AP tracking scenario

In this section, we extend the scenario to analyze the impact of multiple subjects present on the scene, which we tackle using multiple APs. Here, all measurements are performed using AP1 and AP2 in E1. We first consider the results obtained solely by AP1, and then we combine AP1 and AP2. Several experiments are carried out with 2 to 5 subjects, performing different activities. In

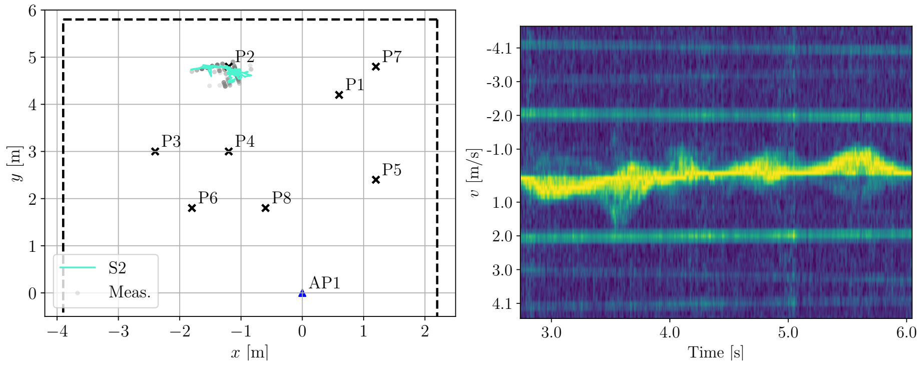


Figure 4.9: Estimated position of a subject sitting down (left) and a portion of the corresponding μD signature (right) extracted by RAPID.

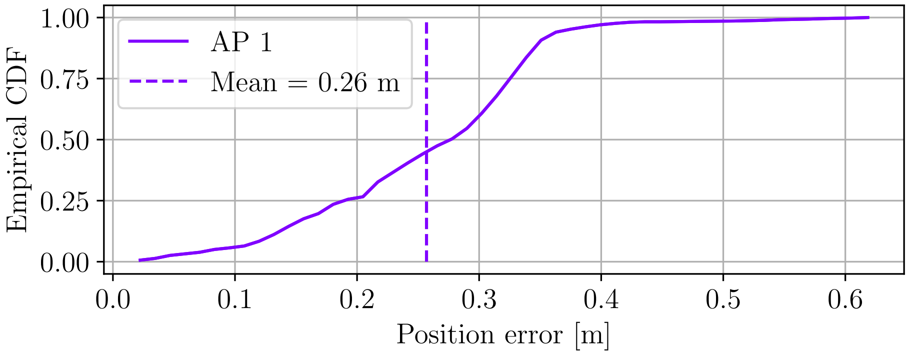


Figure 4.10: Empirical CDF of the positioning error for a subject sitting down in correspondence of marker P2.

total, we collect 28 such sequences each with duration $\sim$ 10 s, of which 13 include 2 subjects, 5 include 3 subjects, 6 include 4 subjects and 4 include 5 subjects.

Presence of multiple subjects. Fig. 4.12 shows some example trajectories estimated by the EKF using the measurements from AP1. RAPID is able to successfully track the users with considerable accuracy in most cases, even for 5 subjects (see Fig. 4.12d). Note that this setup is extremely challenging, especially when more than 3 subjects are present, due to the small dimensions of the environment that lead to a high probability of occlusion happening, i.e., one subject covers the Line-of-Sight (LOS) path between the AP and another individual. mmWave signals do not propagate through the human body, and occlusion may cause missed detection and tracking errors. On the other hand, in real-life scenarios occlusion may happen frequently, and the system must be robust to these events. In Fig. 4.11, we report a quantitative analysis of the effect of increasing the number of subjects in terms of the percentage of subjects that are correctly detected and tracked by RAPID. Using only AP1 we observe that, despite achieving adequate tracking performance, the system capability of detecting the subjects decreases significantly as their number increases. In particular, on average one subject goes undetected when 5 individuals are present.

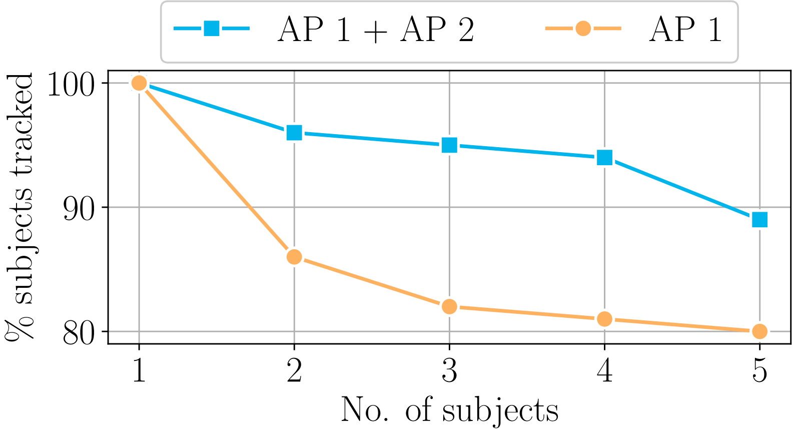


Figure 4.11: Rate of detection with a varying number of subjects using only AP1 and the combination of AP1 and AP2.

Improvement with multiple APs. Combining the FoV of AP1 and AP2 effectively decreases the probability of occlusion events happening, as when the LOS between an AP and a subject is blocked, the other AP can exploit its own LOS path to detect the person. In Fig. 4.13 we report a qualitative example of this, by showing how combining the 2 APs, RAPID can recover from an occlusion event (in this case, 3 subjects are present in the environment). The EKF estimated trajectories from AP1 are shown in Fig. 4.13a: subjects S1 and S2 are successfully detected and tracked, while S3, who is waving hands in P3, is not. This is due to a combination of the occlusion caused by S1 and the fact that P3 is placed at the edge of the FoV of AP1. However, the position of AP2 enables it to detect S3 successfully, while the trajectory of S1 can only be partially reconstructed. Considering the trajectories estimated by both APs, RAPID can detect and track all subjects, successfully extracting their μD signatures, which are reported in Fig. 4.14.

The subject detection rate is also significantly improved by using multiple APs, as shown in the blue curve in Fig. 4.11. Despite AP1 and AP2 being placed along the same axis (x), and only 1.8 m apart, this is sufficient to increase subject detection probability by 11%, 16%, 16% and 11% for the cases of 2, 3, 4 and 5 subjects, respectively.

Finally, we show the impact of *averaging* the positions estimated by the two different APs, see Fig. 4.15. We repeat the experiment described in Section 4.6.2 with a single subject sitting down in position P2. Even using this simple fusion method, RAPID achieves a significant gain in the tail of the localization error distribution. A subject positioned in P2 represents a worst-case for this kind of analysis in our setting, as the locations of the APs with respect to this point are very similar in terms of distance and angle. The same experiment is repeated for position P4, showing a larger improvement from combining the APs. In this case, RAPID goes from an average localization error of 0.35 m using the single APs independently, down to an error of 0.08 m by averaging their estimates. This is due to the more favorable positions from which P4 is illuminated by the APs.

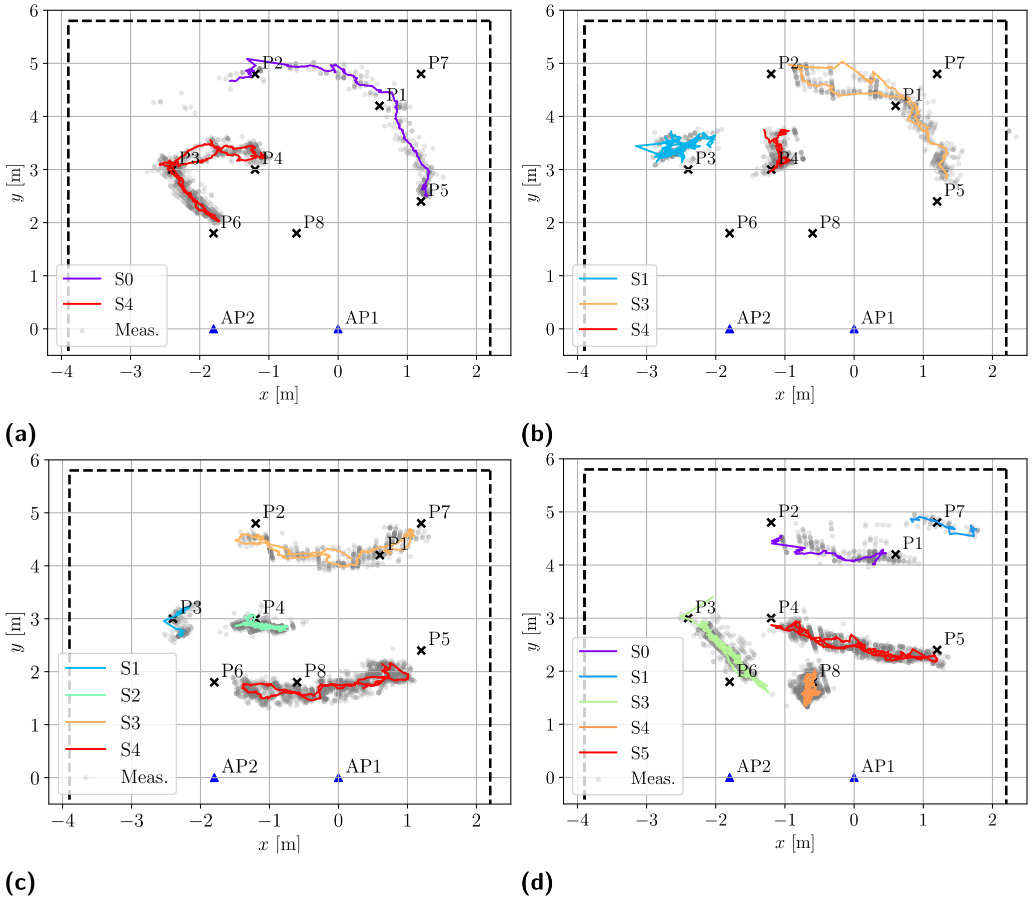
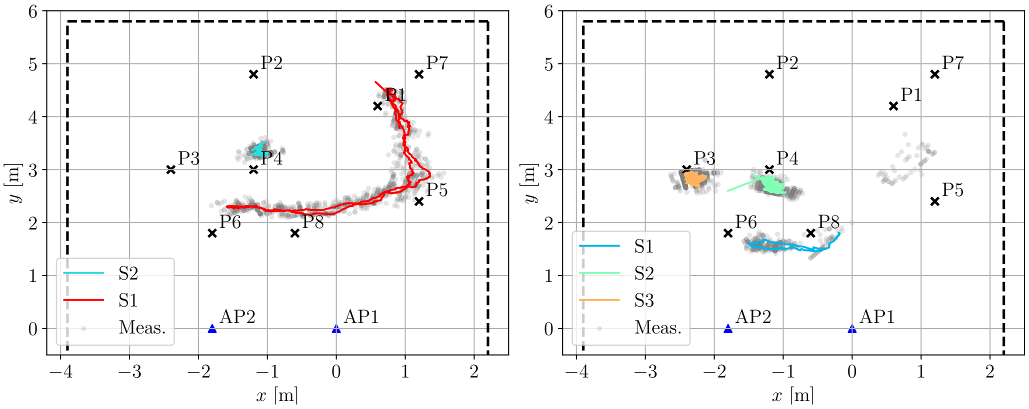


Figure 4.12: EKF trajectories obtained in the multiperson scenario. Here a single AP is used (AP1). We show four successful cases in which RAPID is able to reconstruct the movement trajectories of 2 **(a)**, 3 **(b)**, 4 **(c)** and 5 **(d)** people moving the the room.

4.6.4 Human activity recognition

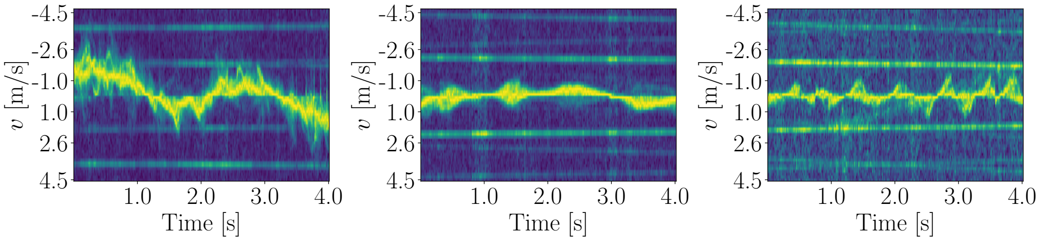
Next, we evaluate the HAR performance of RAPID, comparing it to legacy sub-6 GHz Wi-Fi systems. For all the experiments in this section, unless stated otherwise, we used a unique labeled training dataset of simultaneous IEEE 802.11ay CIR (at 60 GHz) and IEEE 802.11ac Channel Frequency Response (CFR) (at 5 GHz) sequences, which we collected in E1, with a single subject performing the 5 different activities A0 − 4. We used a single RAPID AP and a pair of transmitter/receiver IEEE 802.11ac routers with 4 antenna elements (ASUS RT-AC86U implementing the Nexmon-CSI firmware modifications [118]). The estimates are obtained with the two systems operating *(i) concurrently*, i.e., each training/testing sequence for the same activity of the subject is collected with both the RAPID mmWave AP and the sub-6 GHz system, and *(ii)* with the same μD frequency range and resolution. The latter is achieved by tuning the IEEE 802.11ac system inter-packet transmission time using a slight modification of Eq. (4.14) for the case of non



(a) AP1 estimated trajectories.

(b) AP2 estimated trajectories.

Figure 4.13: Impact of using multiple APs on the occlusion problem. Here, AP1 fails to detect and track S3, while AP2 can only partially reconstruct the trajectory of S1. The combination of the 2 APs successfully detects and tracks all subjects.



(a) S1 running.

(b) S2 sitting down.

(c) S3 waving.

Figure 4.14: Extracted μD signatures of the subjects in Fig. 4.13.

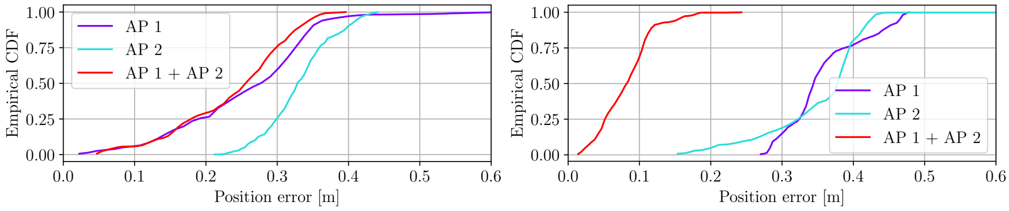


Figure 4.15: Localization error CDFs for a subject sitting dwon in P2 (left) and in P4 (right). Combining multiple APs brings the largest improvement when their point-of-view on the subject is the most diverse.

co-located transmitter and receiver, i.e., $\Delta v = c/(f_o^{\rm ac} M T_c^{\rm ac})$ with $f_o^{\rm ac} = 5$ GHz. Therefore, the IEEE 802.11ac inter-packet transmission time is computed as $T_c^{\rm ac} = 2T_c f_o / f_o^{\rm ac} \approx 6$ ms. The data are obtained in sequences of approximately 10 s, for a total of around 6 minutes of CIR/CFR measurements per activity. Next, the μD spectrograms are obtained from the collected data. To do this in the sub-6 GHz system, we adopt the pre-processing steps proposed in [119], to which we refer for additional details.

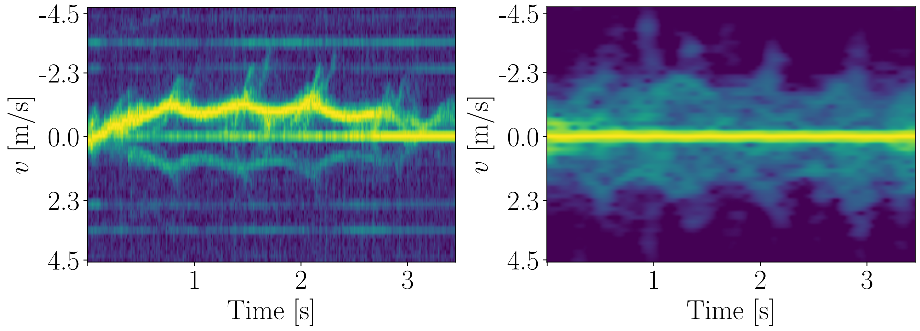


Figure 4.16: Walking spectrogram concurrently obtained with RAPID at 60 GHz (left) and with sub-6 GHz sensing (right).

The resulting μD spectrograms are split into partially overlapping windows of 1.728 s, which are the input to the CNN. For RAPID, we use windows containing $N_{\mu\text{D}} = 200$ time-steps while for the sub-6 GHz setup each window consists of 287 samples. In Fig. 4.16 we show an example of the μD signatures obtained by RAPID and by the sub-6 GHz system for the same measurement sequence of a walking person. We use the CNN model detailed in Section 4.3.5 for both mmWave and sub-6 GHz spectrograms. The CNN is trained using the cross-entropy loss function [53] and the Adam optimizer [120], with learning rate 10^{-4}, until convergence of the loss function on a subset of the training data, used as validation set. We evaluate the performance of the classifier with a weighted average of the per-class F1-score metric, based on the number of samples per class. The F1-score is defined as $\texttt{tp}/[\texttt{tp} + 0.5(\texttt{fp} + \texttt{fn})]$, where $\texttt{tp}$, $\texttt{fp}$ and $\texttt{fn}$ are the predicted true positives, false positives and false negatives, respectively.

Single person, single AP scenario. In Tab. 4.1 we report the confusion matrix and per-class F1-scores obtained by RAPID (grey rows) and by the IEEE 802.11ac system (white rows) on test sequences containing data from the same subject present in the training set, collected in E1. This evaluation is also referred to as our *baseline* HAR experiment in the following. Comparing the two systems, one can see that RAPID accurately classifies all activities, only showing slightly lower performance on A2, sitting down, as this mostly involves body movements directed along an orthogonal direction with respect to the receiver (along the vertical axis). Indeed, the motion-induced μD phase displacement is only measurable in the *radial* direction as we rely on the direct path between the subject and the AP. Sub-6 GHz, instead, benefits from a richer multipath environment and better recognizes A2, but confuses the other activities, especially walking with running and standing still. This is due, in part, to the low resolution of the μD obtained at 5 GHz, which contains coarser-grained information (see Fig. 4.16).

Impact of unknown environment and subject. Next, we further evaluate the HAR robustness of the two systems in more complex settings, involving a different room than the one used for the training data collection (E2), and a different subject performing the activities. Fig. 4.17 reports the weighted average of the per-class F1-scores obtained with RAPID and the sub-6 GHz system: (a) in the baseline scenario, (b) in a different room, E2, on the same subject (c) with a different subject, in the same environment (E1) and (d) in a different environment (E2) and

Table 4.1: Confusion matrix and F1-scores for the baseline case. Grey/white rows refer to RAPID and sub-6 GHz, respectively.

	Predictions [%]				
True [%]	Walking	Running	S. down	Waving	Still
Walking	**97.7**	0	2.3	0	0
	61.4	18.3	0	0	20.3
Running	0	**100**	0	0	0
	0.6	87.1	0	0	12.3
S. down	0	0	95.9	0	4.1
	0	0	**100**	0	0
Waving	0	0	0	**100**	0
	0	0	0.1	85.9	14.0
Still	0	0	0	0	100
	0	0	0	0	100
F1-score [%]	**98.8**	**100**	96.3	**100**	**98.4**
	75.8	91.2	**99.9**	92.4	85.4

on a different subject. The results show that RAPID outperforms the sub-6 GHz counterpart in generalizing to new environments and subjects, showing much lower performance degradation when moving to an unknown room or testing on a different person. In scenario (d) the sub-6 GHz HAR system completely fails, obtaining a very low F1-score, due to the challenging combination of a different room and a different subject. Conversely, RAPID still achieves good performance. We stress that here the training data contain measurements from only one subject. Therefore, the CNN classifier must possess great generalization capabilities to correctly classify the activities performed by another person, as they may have slightly different features.

In addition, we test the two systems under *interference* from another subject in one of the activities of the training set, as shown in Tab. 4.2. For this, we use the same setting as in the baseline, but we replace the training data for A3, waving hands, with new measurements where another person, termed *interfering subject*, is present in the room besides the subject performing A3. The interfering subject performs a different, randomly selected, activity in each measurement sequence, in a position close to the intended subject, thus possibly disturbing the useful signal reflections. RAPID, thanks to the separation between different subjects enabled by the high ranging accuracy of mmWaves and the tracking process, is highly robust to the presence of other people. Sub-6 GHz sensing, instead, suffers from its low ranging resolution ($\sim$ 4 m) and is greatly affected by the interference.

Multi-person, multi-AP scenario. Next, we evaluate RAPID's HAR performance degradation when multiple subjects are concurrently present in the environment, each performing, in general, a different activity. The aim here is to assess the effectiveness of RAPID in the separation of μD signatures associated with different targets. In this evaluation, we do not consider the sub-6 GHz

Table 4.2: HAR performance under interference from another subject in the training dataset.

F1-score [%]	Walking	Running	S. down	Waving	Still
RAPID	**98.5**	**99.9**	93.2	**100**	**96.6**
Sub-6 GHz	72.2	92.4	**97.8**	58.0	77.8

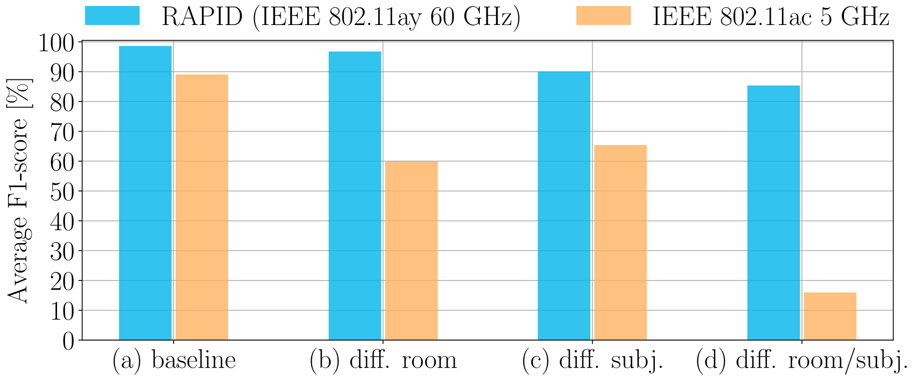


Figure 4.17: Comparison between the HAR F1-score obtained by RAPID and by standard IEEE 802.11ac sensing at 5 GHz for various scenarios.

system, as the intrinsic limits in terms of ranging (~ 4 m) and angular ($\sim 20°$) resolutions prevent people tracking in crowded indoor scenarios such as the ones under study [100], thus making the separation of the multiple subjects infeasible.

We collect a labeled training dataset including 6 subjects performing the 5 different activities $A0-4$ using a single RAPID-AP. The data are obtained in sequences of approximately 10 s, and the resulting μD spectrograms are split into windows of 1.728 s as in the single target case. In total, this dataset contains around 2 minutes per activity *per subject*. By training on different subjects, we aim at mitigating the HAR performance reduction due to the difficulty of generalizing to different people, to better gauge the sole effect of μD separation. We test the trained model on the same multi-person sequences used in Section 4.6.3, adding 6 additional sequences with a single subject, for a total of 34 sequences. We use the RAPID processing steps to extract the μD signatures of each subject's movement; when using 2 APs, we use the decision fusion scheme from Section 4.3.5.

Tab. 4.3 shows the F1-score of RAPID for a varying number of people in the scene, and the gain obtained by combining the 2 APs with respect to using only AP1. In addition, we also report the corresponding detection rate, previously shown in Fig. 4.11, for completeness. We observe that the F1-score only slightly decreases when moving from 2 to 5 subjects. This shows that the proposed μD extraction process can reliably separate the contributions of the different individuals. In addition, combining multiple APs can bring a slight improvement in some cases, by exploiting the different illumination angles of the devices.

Table 4.3: HAR F1-score and detection rate vs. no. of concurrent users.

APs	Metric	1 subj.	2 subj.	3 subj.	4 subj.	5 subj.
1	F1	99.9	99.3	97.9	95.3	94.4
	Det. rate	100	86.1	82.9	81.3	80.0
1 & 2	F1	100	99.4	99.4	95.4	94.4
	Det. rate	100	96.7	95.5	94.5	89.2

4.6.5 Person identification

In this section we test the performance of RAPID on person identification, by building a dataset including the gait μD spectrograms of 7 subjects, collected in E1. We collect from 3 to 5 minutes of training data per subject. The input samples for the classifier are obtained using μD windows of the same length as for HAR, i.e., 1.728 s. The CNN classifier is trained using the same parameters and loss function used for HAR.

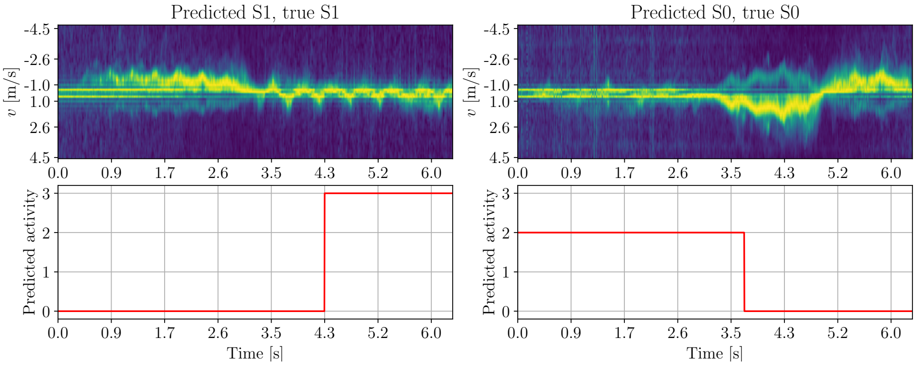
Person identification accuracy. First, we evaluate the accuracy of person identification on a varying number of subjects to recognize. In Tab. 4.4 we report the accuracy values obtained by RAPID when increasing the number of subjects from 2 to 7. The obtained values are not significantly lower from those obtained with mmWave radars, and in some cases even superior, e.g., the 79% on 5 subjects in [7], the 98% with 4 subjects in [5] or the 89% with 12 subjects in [8]. This is even more valuable considering the few available training data and the short duration of the observation window used, compared to the windows used in the mentioned papers which vary between 2 and 3 s.

Continuous HAR and person identification. Finally, we show that RAPID is capable of simultaneously *(i)* tracking subjects, *(ii)* recognizing their activities, and *(iii)* identifying who is performing each activity from their gait. We perform several tests in which 2 subjects, *concurrently* present in the room, perform various activities sequentially, e.g., walking then sitting, etc. In this scenario, people tracking is of key importance to collect the temporal evolution of each subject's μD, so that all the activities performed by a person can be associated to that person's identity, obtained by RAPID when he/she is walking.

In Fig. 4.18 we show the results obtained by RAPID with 2 subjects, S0 and S1, behaving as follows. S0 enters the scene walking, then after approximately 3.5 s S0 stops and starts waving hands, while S2 is sitting down and then starts walking after 3.5 s. We report the μD signature extracted after successfully tracking the subjects, along with the predicted activity using our moving window approach. We observe that RAPID detects the change in the activity performed by each subject; moreover, by applying the identification CNN to the spectrogram portion where the subjects are walking, it successfully identifies them as S0 and S1 among the 7 subjects in the training set.

Table 4.4: Identification accuracy vs. number of subjects.

	2 subj.	3 subj.	4 subj.	5 subj.	6 subj.	7 subj.
Acc. [%]	97.8	95.9	94.6	94.1	92.7	90.0



(a) S1 walking-waving.

(b) S0 sitting down-walking.

Figure 4.18: μD signature and corresponding CNN output when subject 0 is sitting down (A2), then starts walking (A0), while subject 1 is walking and then starts waving hands (A3).

4.6.6 Overhead considerations

The sensing operations performed by RAPID add a certain level of overhead to the communication process, due to appending TRN units to the communication packets. We can asses the overhead of RAPID by comparing the PHY layer packet size in IEEE 802.11ay to the size of TRN fields used for sensing. As shown in Fig. 4.4, physical layer Protocol Data Unit (PDU) include the Short Training Field (STF), the CEF and the PHY layer header, including $\mathrm{STF}_l = 2176$, $\mathrm{CEF}_l = 1152$ and $\mathrm{PHY}_l = 1024$ samples, respectively [111]. Each TRN field includes 6 complementary Golay sequences, for a total of $\mathrm{TRN}_l = 768$ samples. Therefore, the overhead introduced by appending ξ TRN fields to a packet is

$$O = \frac{\mathrm{TRN}_l \cdot \xi}{\mathrm{STF}_l + \mathrm{CEF}_l + \mathrm{PHY}_l + \mathrm{DATA}_l + \mathrm{TRN}_l \cdot \xi}, \tag{4.19}$$

where DATA_l is the length of the data portion of the packet. We recall that, with RAPID, it is sufficient to illuminate a person with *one* BP to apply the extraction of the μD spectrum, and that we can use one BP per TRN field, so ξ can be selected equal to the number of subjects tracked by RAPID. In order to reduce the inefficiency of the MAC layer and achieve Gigabit data rates, in IEEE 802.11ay large packet aggregation is permitted, allowing PHY layer PDU to contain up to 4 MB of data. For this, multiple MAC layer PDU of 1.5 kB are encapsulated into a single PHY layer packet. Compared to these large packet sizes, the TRN fields used by RAPID add a limited amount of overhead. To see this, consider that, e.g. Modulation and Coding Scheme (MCS) 8 is used, and that the data size is 20 kB. We get $\mathrm{DATA}_l = 126784$ samples (due to the MCS used)

[111], leading to $O = 0.6 \cdot \xi\%$.

4.7 Concluding remarks

In this chapter, we have designed and implemented RAPID, the first mmWave ISAC system performing high resolution sensing of human μD signatures through standard-compliant IEEE 802.11ay packets. RAPID uses the in-packet TRN fields, as specified by the 802.11ay standard, to estimate the channel impulse response. This makes it possible to perform joint tracking and localization of multiple people freely moving in an indoor environment. In addition, their μD signatures are extracted by analyzing the phase difference between subsequent packets, which allows RAPID to perform advanced sensing tasks such as continuous HAR and person identification, with radar-level accuracy. RAPID successfully combines the high resolution sensing capabilities of mmWave radars with the scalability and ease of deployment of existing communication hardware, allowing the seamless integration of multiple APs. We implemented two RAPID APs with full-duplex capabilities on an FPGA-based SDR platform equipped with phased antenna arrays, and we have thoroughly evaluated the system performance through an extensive measurement campaign. Our results show that 2 combined RAPID-APs can track up to 5 subjects concurrently moving in an indoor environment, achieving accuracies of up to 94% and 90% for HAR and person identification, respectively. Moreover, in HAR, RAPID performs significantly better than standard sub-6 GHz sensing, showing better capability of distinguishing similar activities and generalizing to new environments and unkwnown subjects.

5
A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems

5.1 Introduction

In this chapter, we address the problem of enabling ISAC in realistic mmWave communication systems. Our aim is to reuse existing communication traffic for sensing as much as possible, thus introducing only a minimal amount of additional overhead. To this end, we propose SPARCS, the first mmWave ISAC system that reconstructs human μD signatures from *irregular and sparse* CIR samples obtained from realistic traffic patterns. The main insight of SPARCS is to leverage the intrinsic sparsity of the reflections in the mmWave channel to pose the μD reconstruction as a sparse recovery problem. Indeed, mmWave CIR estimation can naturally separate signal propagation paths with $<$ 10 cm resolution, leading to a sparse multi-path environment and consequently a sparse CIR in the Doppler domain. This allows obtaining highly accurate μD signatures from only a small, randomly distributed fraction of the CIR samples that are currently needed by existing ISAC methods. To do so, SPARCS first performs CIR resampling to construct a regular grid of CIR samples with missing vales due to the irregularity of the sampling process in time. Next, a sparse reconstruction method is used to obtain the μD spectrum, decoupling different propagation paths to leverage their sparsity property. Lastly, whenever communication traffic is absent or insufficient for the μD extraction, SPARCS supports a dynamic injection of very short CIR estimation fields into the (idle) channel. Given its sparse recovery capabilities, only a few additional CIR sensing units are needed to retrieve the μD, thus entailing a negligible overhead to the communication rate.

SPARCS is compatible with any mmWave system that supports transmit beamforming for directional communication and CIR estimation. This is the case, for example, for IEEE 802.11ay WLANs at 60 GHz, which provide in-packet CIR estimation for beam tracking purposes, and for 3GPP 5G-NR, where base stations can send frequent downlink CSI-Reference Signal (CSI-RS) to estimate the channel using different BPs.

To evaluate SPARCS' performance, we implement it on a 60 GHz IEEE 802.11ay SDR experimentation platform, the same used in Chapter 4. We then test it on sparse and irregular CIR samples derived from standard-compliant traces, both for synthetic traffic and traffic patterns obtained from datasets of operational real-world Wi-Fi APs deployments [121]. To assess the quality of the reconstructed μD signatures, we use them as input for a typical downstream task such as HAR, which classifies human movement detected by the captured μD into different possible activities. The main contributions of this chapter are summarized next.

1. We propose SPARCS, an ISAC method for mmWave systems that can reconstruct high-quality μD signatures of human movement from irregular and sparse CIR estimation samples. SPARCS reuses training fields appended to communication packets as sensing units, and injects additional sensing units if necessary, adapting to the underlying communication traffic and minimizing the sensing overhead.

2. We provide an original formulation of the μD extraction in communication systems as a sparse recovery problem, leveraging the intrinsic high distance resolution and sparsity properties of the mmWave channel. As a side effect, this also improves the quality of the resulting spectrograms, making them more robust to noise and interference.

3. We design and validate an algorithm to perform the injection of additional sensing units when communication traffic is insufficient. The process is dynamic, requires no knowledge about future packet transmissions, and incurs minimal overall overhead.

4. We evaluate SPARCS by implementing it on an IEEE 802.11ay-compliant 60 GHz SDR platform and testing it on CIR measurements collected with realistic Wi-Fi traffic patterns. For the common HAR task, the μD signatures reconstructed by SPARCS achieve better F1 scores than existing methods, while reducing sensing overhead by a factor of 7.

The chapter is organized as follows. In Section 5.2 we recall some key concepts from Chapter 4 regarding mmWave human μD sensing using CIR, discussing their generalization to a generic mmWave system. SPARCS is introduced and explained in detail in Section 5.3, describing the sparse recovery problem formulation and the involved processing steps. In Section 5.4 we discuss the implementation of SPARCS on an SDR platform, and Section 5.5 provides an evaluation of the system on real measurement traces, comparing it to RAPID (see Chapter 4). We summarize the related work in Section 5.6 and give concluding remarks in Section 5.7.

5.2 Primer on mmWave sensing

In this section we give a brief description of the CIR model for mmWave communication systems that we use for sensing. We then describe a baseline approach that allows tracking the movement of people in the environment and extract their μD signatures using *regularly sampled* CIR information. This forms the basis of our SPARCS design, which entirely *eliminates* the requirement of fixed Inter-Frame Spacing (IFS) and enables ultra low-overhead ISAC.

5.2.1 Sensing in mmWave systems

Capturing the movement features of humans in the environment requires an analysis of the reflections of the transmitted signal from their bodies, which is usually carried out applying signal processing techniques to the CIR. Due to the high path loss occurring at mmWave frequencies, directional communication is employed by means of transmitter and receiver beamforming, typically using phased antenna arrays. The transmitter and the receiver use suitable BP configurations of their antenna arrays to maximize the signal strength [111], [117], [122]. To successfully sense with a mmWave system, *at least* one of the BPs has to illuminate the subjects of interest, as only in this case the reflected signal carries detectable information about the movement signature. To this end, similarly to what we did in Chapter 4, we consider a setup where an AP transmits packets and is able to collect the reflections of *its own* signal, after being reflected by objects (including humans). This reflection is collected by the receiver array of the AP itself using a quasi-omnidirectional BP. This requires full-duplex capabilities, as is common in ISAC scenarios [123], which in the simplest form can be achieved with a MIMO system in a mixed configuration with one RF chain as transmitter and another as receiver. The CIR estimation fields used for sensing, which we denote by *sensing units*, can either be piggybacked by appending them as a trailer to the PHY communication packets or transmitted independently (*injected*). mmWave standards implement beam training mechanisms that help to establish a communication link by testing different BP combinations and then selecting the best one. Such functionality is supported by all mmWave standards. For example, 5G-NR [122], use Synchronization Signal Block (SSB) and CSI-RS for beam management, while WLAN systems adopting the IEEE 802.11ad/ay standards [111], [117] use channel estimation and training fields (CEF and TRN, respectively) to obtain accurate CIR information. Our framework to extract sensing information from CIR measurements can be applied regardless of the specifics of the standards.

5.2.2 mmWave CIR model

Due to the large transmission bandwidth of mmWave systems, channel measurements contain fine-grained information about the environment [9], [28], [31]. Depending on the communication system we consider, sensing could be performed using the 5G-NR OFDM Channel State Information (CSI), which contains the channel gains for each OFDM subcarrier, or the IEEE 802.11ad/ay Single Carrier (SC) CIR. Both communication schemes are suitable for human sensing: *(i)* in 5G-NR, the base stations can send frequent downlink CSI-RS to estimate the channel using different

BPs, while *(ii)* in IEEE 802.11ay in-packet beam tracking is enabled, so that specific fields called training fields (TRN), each using a different BP, can be appended to communication packets. In the following, we focus on SC CIR, and show how to extract the μD effect of human movement. However, previous works have demonstrated that similar processing can be performed with OFDM CSI [9], [98], and SPARCS is general enough to be applied in both cases.

As the SC CIR model, we use the one presented in Chapter 4, in Eq. (4.1). However, differently from what we did in the previous chapter, here we do not simplify Eq. (4.1) considering a single propagation path per CIR bin, but we rather address the possibility of having the contributions of $P_\ell(t)$ paths overlapping in bin ℓ (see Section 5.3.2).

5.2.3 micro-Doppler extraction

The extraction of the μD spectrum from multiple, concurrently moving subjects requires tracking the position of each person in the physical space, in order to separate their individual contributions to the CIR. For this we employ the methods described in the previous chapter, specifically in Section 4.3.3, Section 4.3.3, and Section 4.3.3.

Then, spectral analysis over different CIR samples, as described in Section 4.3.4 yields the desired μD signature [31], [100]. As shown in Chapter 4, one of the most computationally efficient methods to perform such spectral analysis is to apply a STFT to the CIR along the slow-time dimension. This kind of processing requires a window of W subsequent estimates of the CIR with a *fixed* CIR sampling interval of T_c seconds, provided that the time spanned by the window is short enough to consider the movement velocity of the reflectors *constant* for its whole duration. Note that this operation allows detecting and separating the velocities of the $P_\ell(t)$ reflectors, whose contributions overlap in path ℓ when considering a single estimate of the CIR. The choice of T_c impacts the frequency resolution of the STFT, $\Delta f^d = 1/(WT_c)$, and its maximum measurable frequency, $f^d_{\max} = 1/(2T_c)$. Using the relationship between the Doppler frequency and the corresponding velocity, one can obtain the velocity resolution and the maximum observable velocity as $\Delta v = c/(2f_oWT_c)$ and $v_{\max} = c/(4f_oT_c)$. To fully capture the range of velocities of interest for human movement, the typical approach is to select T_c such that $v_{\max}$ is sufficiently high that is covers the velocities that can occur in the human activities of interest, which may vary depending on the application [4], [7], [31].

In Chapter 4 we assumed that the constraint of a fixed T_c is met, which does not hold in realistic communication scenarios, where packet transmissions are scheduled according to the needs of the communication protocols rather than sensing accuracy. Traffic patterns are typically bursty and irregular and thus cannot be used by existing methods for human sensing. Instead, dedicated time slots need to be reserved for the transmission of sensing units, which is incompatible with the random access CSMA/CA MAC commonly used in IEEE 802.11. Conversely, SPARCS is the first approach that does not require any specific pattern in the transmission of the sensing units, enabling true ISAC by exploiting communication packets for sensing whenever possible, and introducing minimal additional overhead when necessary.

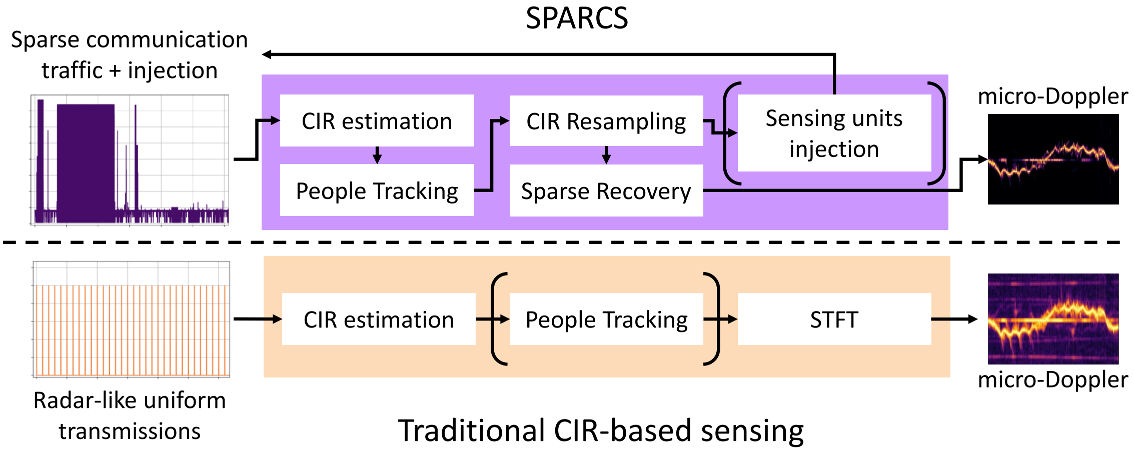


Figure 5.1: Comparison between the traditional CIR-based human sensing and SPARCS.

5.3 SPARCS methodology

We now present the SPARCS algorithm to recover the μD spectrum from irregular and sparse CIR sampling patterns. The processing steps of SPARCS compared to traditional CIR-based sensing methods are shown in Fig. 5.1.

(1) **CIR resampling:** after CIR estimation and people tracking, for which we adopt the standard Joint Probabilistic Data Association Filter (JPDAF) technique [89], we apply a resampling strategy to approximate the irregularly spaced CIR values with a regular sequence whose sampling interval is chosen according to the desired μD resolution (Section 5.3.1). Due to irregularity of the original sampling process, the approximated regular sequence may contain missing values that need to be *filled* in the subsequent processing steps.

(2) **Sparse μD recovery:** we formulate the recovery of the μD spectrum from the incomplete CIR measurements as a sparse recovery problem. For this, we leverage two key aspects. On the one hand, the intrinsic sparsity of the mmWave channel leads to a few signal reflections from the human body that carry information about different body parts. On the other hand, the high distance resolution of mmWave systems makes the reflections from the different body parts separable. The combined effect of these two properties is that the resulting CIR is highly sparse in the Doppler frequency domain, as detailed in Section 5.3.2 We then solve the sparse recovery problem using the Iterative Hard Thresholding (IHT) algorithm for each CIR path (Section 5.3.3), and aggregate the results to obtain the final μD spectrum (Section 5.3.4).

(3) **Sensing unit injection:** when communication traffic is absent or too scarce to obtain an accurate reconstruction, our system can inject short sensing units into the (idle) channel to overcome the problem, as described in Section 5.3.5. Thanks to the sparse reconstruction of point (2), the amount of units that need to be injected is minimal and can be tuned to trade off between overhead and sensing accuracy.

5.3.1 CIR resampling

Our system samples the CIR at time instants t_i, which coincide with the reception of the reflections from the i-th transmitted packet. To reconstruct the μD spectrum from CIR samples which are randomly distributed in the time domain, we first resample the CIR to obtain regularly spaced samples with a fixed granularity T_c, where possible. To do so, we resort to the *slotted resampling* technique, which allows approximating a sequence of randomly spaced samples into a regular grid with *missing values* [124]. We consider N_s consecutive samples obtained at the time instants $t_0, t_1, \dots, t_{N_s}$ and denote by $0, T_c, 2T_c, \dots, (K-1)T_c$ the regular grid with step size T_c. Slotted resampling constructs a new CIR sample sequence $h_{\ell,b}(kT_c)$ where the CIR values are obtained from the original sequence $h_{\ell,b}(t_i)$ as follows. Time *bins* (or intervals) of length T_c are centered on each time instant of the regular grid, i.e., bin k is $\beta_k = [kT_c - T_c/2, kT_c + T_c/2)$, with center kT_c. Then, the value of the CIR corresponding to the k-th grid value is either *(i)* selected among the values of the original sequence whose sampling times fall inside bin k, taking the one whose sampling time is the closest to the bin center, or *(ii)* considered as a missing value if no samples of the original sequence fall inside bin k. Specifically,

$$h_{\ell,b}(kT_c) = \begin{cases} 0 & \text{if } \{t_i | t_i \in \beta_k\} = \emptyset, \\ h_{\ell,b}(t_k) & \text{otherwise,} \end{cases} \tag{5.1}$$

where the 0 values represent missing samples and

$$t_k = \underset{\tau \in \{t_i | t_i \in \beta_k\}}{\operatorname{argmin}} |kT_c - \tau|. \tag{5.2}$$

The resulting, regularly spaced sequence of CIR samples is used to reconstruct the μD spectrum of the subject. However, due to the missing samples which are set to 0, a plain application of the STFT (as described in Section 5.2.3) would lead to a corrupted spectrum. In the next section we detail our solution to this problem, which is based on sparse recovery techniques.

5.3.2 Sparse μD recovery problem formulation

Several methods exist to tackle the problem of computing the power spectrum of non-uniformly sampled signals [124]. Our approach belongs to the category of sparsity-based approaches, in which the sparsity of the signal in the frequency domain is leveraged to drastically reduce the number of measurements needed for an accurate reconstruction of the spectrum. We select windows of length W samples (window size) every δ samples from the sequence $h_{\ell,b}(kT_c)$, choosing $\delta = W/2$. In the following we consider $W = N_D$, which means that the number of frequency components that we extract from the signal is the same as the length of the time window, i.e., we do not apply padding. Due to the slotted resampling process, each window may contain missing samples. We denote by $\mathcal{U}_m$ the set of indices of the available samples contained in the m-th window. Then, we define vector $\mathbf{h}_{\ell,b}(m) \in \mathbb{C}^{|\mathcal{U}_m|}$, containing the available CIR samples in the m-th window, and vector $\tilde{\mathbf{h}}_{\ell,b}(m) \in \mathbb{C}^W$, representing the complete m-th CIR window, which is only partially known due to the missing samples. We also denote by $\mathbf{F}_{\text{inv}}$ the inverse Fourier matrix, whose element in position (g, l) is given by $(\mathbf{F}_{\text{inv}})_{gl} = (1/\sqrt{W}) \exp(j2\pi gl/W)$, $g, l = 0, \dots, W-1$ while $\mathbf{U}_m = \left[\mathbf{u}_i^T\right], \forall i \in \mathcal{U}_m$ is the matrix that selects the rows of $\mathbf{F}_{\text{inv}}$ whose indices are in $\mathcal{U}_m$. $\mathbf{u}_i$ is the vector of all zeros but the i-th component, which equals 1.

The following relation holds between the incomplete CIR window, $\mathbf{h}_{\ell,b}(m)$, and the Fourier Transform (FT) of the full CIR window, $\mathbf{H}_{\ell,b}(m) \in \mathbb{C}^W$, which we aim to recover in order to compute the μD spectrum,

$$\mathbf{h}_{\ell,b}(m) = \mathbf{U}_m \tilde{\mathbf{h}}_{\ell,b}(m) = \mathbf{U}_m \mathbf{F}_{\text{inv}} \mathbf{H}_{\ell,b}(m) = \mathbf{\Psi}_m \mathbf{H}_{\ell,b}(m), \tag{5.3}$$

where in the last step we use matrix $\mathbf{\Psi}_m = \mathbf{U}_m \mathbf{F}_{\text{inv}}$ as a shorthand notation. Given Eq. (5.3), our aim is to recover $\mathbf{H}_{\ell,b}(m)$ from the incomplete measurement vector $\mathbf{h}_{\ell,b}(m)$, which is a typical sparse recovery or compressed sensing problem [29]. In this framework, it has been proven that recovering the FT of the desired signal is possible if the latter is sparse in the frequency domain, i.e., the FT only contains a low fraction of non-zero elements. To verify that this sparsity assumption holds in our case, we rewrite Eq. (4.1) after the resampling and windowing operations, so that the i-th sample of the complete m-th window is given by

$$\left[\tilde{\mathbf{h}}_{\ell,b}(m)\right]_i = \sum_{p=1}^{P_\ell(m)} a_{\ell,b}^p(m) \exp\left\{-j4\pi \frac{f_o}{c}\left[d_\ell^p + (m\delta + i)T_c v_{\ell,m}^p\right]\right\}, \tag{5.4}$$

where $v_{\ell,m}^p$ is the radial velocity of the p-th reflector in path ℓ during window m, and d_ℓ^p its distance from the AP. Here, we use the assumption from Section 5.2.3 that the velocity of each reflector can be considered constant during a window. In addition, we also consider that the reflective coefficients and the number of reflectors are constant. This is reasonable for the considered setup, where the reflectors are parts of the human body, which typically move slowly compared to the duration of a window WT_c (see also Section 5.4).

From Eq. (5.4), one can see that as long as $P_\ell(m) \ll W$, the FT of $\tilde{\mathbf{h}}_{\ell,b}(m)$ is indeed sparse, as it is composed of $P_\ell(m)$ spectral lines located at frequencies $2f_o v_{\ell,m}^p/c$. Given the excellent distance resolution due to the high bandwidth of mmWave systems and the *intrinsic* sparsity and directionality of the channel, the different parts of the subject's body tend to contribute to the μD spectrum in different CIR paths as shown in Fig. 5.2. Therefore, $P_\ell(m)$ is generally close, if not equal, to 1. Sometimes the number of reflectors in a single path can be larger than 1, due to different body parts being closer than the distance resolution of the system, but this number is still much lower than W. This even holds for multiple subjects. Assume that two subjects with labels 1 and 2 are present in the monitored physical space, and denote by (ℓ_1, b_1) and (ℓ_2, b_2) their CIR path-BP pairs. According to Eq. (5.4), the sparsity assumption must hold for each pair *independently*, and this is verified as long as the subjects occupy different spatial positions. Specifically, *(i)* if $\ell_1 \neq \ell_2$ the CIRs along BPs b_1 and b_2 are the combination of $P_{\ell_1}(m) \ll W$ and $P_{\ell_2}(m) \ll W$ complex exponentials each, and *(ii)* if $\ell_1 = \ell_2$, but $b_1 \neq b_2$, the attenuation coefficient of b_1 will mostly remove the reflection from subject 2 in $\tilde{\mathbf{h}}_{\ell_1,b_1}$ and vice versa, making

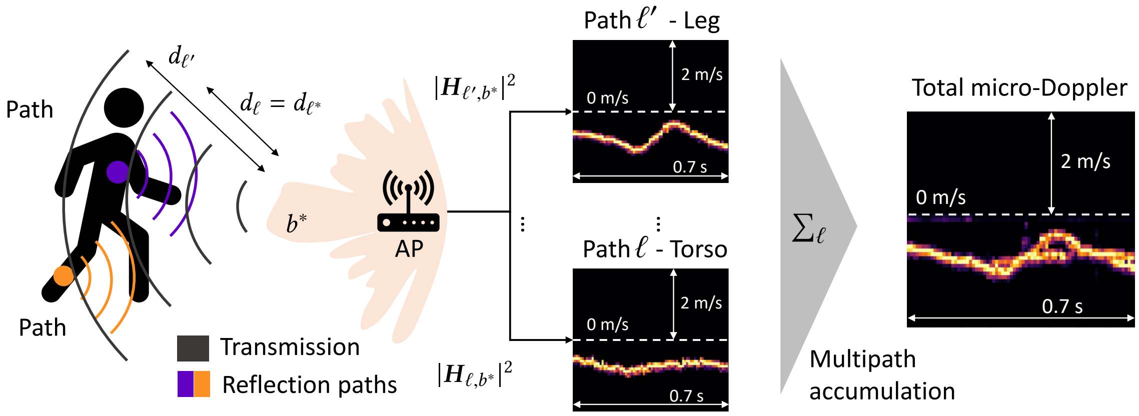


Figure 5.2: Visual representation of the μD spectrum computed using SPARCS on 2 different CIR paths, one containing the reflection from a person's torso, the other capturing the μD signature of the leg. The total μD is obtained summing together these contributions.

Algorithm 5.1 Single path sparse recovery.

```
Input: $\mathbf{h}_{\ell,b^*}(m)$, $\eta, n_{\max}, \Omega, \xi$.
Output: $\mathbf{H}_{\ell,b^*}(m)$.
Collect the set of available samples indices $\mathcal{U}_m$.
Build matrices $\mathbf{U}_m = \left[\mathbf{u}_i^T\right], \forall i \in \mathcal{U}_m$ and $\mathcal{F}_{\text{inv}}$
Compute $\mathbf{\Psi}_m = \mathbf{U}_m \mathcal{F}_{\text{inv}}$.
Set $\hat{\mathbf{H}}^{(0)} = \mathbf{0}$, $n = 0$, $\gamma^{(0)}$ to any value $> \xi$.
while $n < n_{\max}$ and $\gamma^{(n)} > \xi$
    $\hat{\mathbf{H}}^{(n+1)} \leftarrow$ Eq. (5.7)
    $\gamma^{(n+1)} \leftarrow ||\hat{\mathbf{H}}^{(n+1)} - \hat{\mathbf{H}}^{(n)}||_2$
    $n \leftarrow n + 1$
end while
return $\hat{\mathbf{H}}^{(n)}$
```

the contributions from the subjects separable. The contributions from different subjects overlap only if they occupy the same CIR path *and* share the same BP, which is very unlikely to occur in real cases due to the high distance (~ 8 cm) and angular (as low as 2°) resolutions of the mmWave CIR [31]. Therefore, the sparsity assumption in SPARCS still holds even if multiple subjects are present in the environment. Due to this, we can assume that $P_\ell(m) \ll W$ holds, and that sparse recovery techniques can be used to recover $\mathbf{H}_{\ell,b}(m)$, as detailed in the next section.

5.3.3 Single-path sparse recovery

Given the model from Eq. (5.3), the reconstruction of the CIR FT along each path can be posed as a sparse recovery problem. Specifically, we seek a vector $\mathbf{H}_{\ell,b}(m)$ which is a solution to Eq. (5.3) while being as sparse as possible, coherent with the above discussion. Considering the BP b^* pointing in the direction of the target, the desired FT of $\tilde{\mathbf{h}}_{\ell,b^*}(m)$ is the solution of the optimization problem

$$\mathbf{H}_{\ell,b^*}(m) = \underset{\mathbf{H}}{\mathrm{argmin}} ||\mathbf{H}||_0 \quad \text{s.t.} \ ||\mathbf{h}_{\ell,b^*}(m) - \mathbf{\Psi}_m\mathbf{H}||_2 \leq \varepsilon, \tag{5.5}$$

where $||\cdot||_0$ denotes the ℓ_0-norm of a vector, i.e., the number of its non-zero components. The constant $\varepsilon > 0$ can be estimated from the noise in the CIR, using a training dataset.

An approximate local solution to Eq. (5.5) can be found using fast greedy algorithms [29]. We adopt the IHT, which solves

$$\mathbf{H}_{\ell,b^*}(m) = \underset{\mathbf{H}}{\mathrm{argmin}} ||\mathbf{h}_{\ell,b^*}(m) - \mathbf{\Psi}_m\mathbf{H}||_2^2 \quad \text{s.t.} \ ||\mathbf{H}||_0 \leq \Omega, \tag{5.6}$$

where Ω is a pre-defined sparsity level parameter. The algorithm involves an iterative gradient descent step on the quadratic term in Eq. (5.6), followed by a thresholding operation:

$$\hat{\mathbf{H}}^{(n+1)} \leftarrow \mathcal{T}_\Omega \left[\hat{\mathbf{H}}^{(n)} + \eta \mathbf{\Psi}_m^T \left(\mathbf{h}_{\ell,b^*}(m) - \mathbf{\Psi}_m \hat{\mathbf{H}}^{(n)}\right)\right], \tag{5.7}$$

where n is the iteration index and $\mathcal{T}_\Omega$ is the hard-thresholding operator, which sets to 0 all the components of the argument vector except the Ω largest ones in terms of the Euclidean norm. η is a learning rate parameter which can be tuned to improve the convergence properties. The iterative process is stopped whenever $||\hat{\mathbf{H}}^{(n+1)} - \hat{\mathbf{H}}^{(n)}||_2 < \xi$ or when a maximum number of iterations, $n_{\max}$, is reached. In SPARCS, Ω is a key parameter, which is strictly related to the number of reflectors $P_\ell(m)$: as IHT reconstructs a vector which has at most Ω non-zero elements, Ω is an upper bound for $P_\ell(m)$, and it can be thought of as the maximum number of reflectors per path that we allow reconstructing. Ω can be tuned in order to obtain better μD reconstruction (see Section 5.5.5). The sparse recovery algorithm is summarized in Alg. 5.1. According to the compressive sensing theory [125], the reconstruction performance of IHT (and in general of any recovery algorithm) degrades as the number of available measurements, $|\mathcal{U}_m|$, decreases. Theoretical results show that the minimum number of measurements needed to reconstruct $\mathbf{H}_{\ell,b^*}(m)$ is $\mathcal{O}(\Omega \log(W/\Omega))$ [125], although the exact number has to be estimated empirically as it also depends on the level of noise present in the signal. In Section 5.5, we show that SPARCS can achieve excellent μD reconstruction with as few as $W/8$ measurements per window, thanks to the high sparsity of the mmWave CIR.

5.3.4 Multi-path aggregation

The moving body of a person causes several reflections that affect more than one CIR path, as discussed in Section 5.3.2. Using the procedure described in the previous sections, SPARCS is able to retrieve the contribution of each path $\mathbf{H}_{\ell,b^*}(m)$ to the μD. Since the different body parts contribute to the μD in different paths, to fully capture human movement we need to combine the information from the different paths. Denote by Q the number of distance bins we aggregate to obtain the μD spectrum. For convenience, we assume Q to be an odd integer, as this makes the following processing steps symmetric with respect to a central CIR path (corresponding to the torso), but the same steps can be applied for Q being even. We aggregate the spectra obtained from the path caused by the torso, ℓ^*, with the $\lfloor Q/2 \rfloor$ distance bins preceding ℓ^* and the $\lfloor Q/2 \rfloor$ subsequent distance bins, as they may contain the contributions of the other body parts. Using vector notation, the expression of the total μD spectrum in Eq. (4.12) is

$$\mathbf{D}(m) = \sum_{\ell=\ell^*-\lfloor Q/2 \rfloor}^{\ell^*+\lfloor Q/2 \rfloor} \left| \mathbf{H}_{\ell,b^*}(m) \right|^2 , \tag{5.8}$$

where the squared magnitude is applied element-wise. We then apply the same min-max normalization in Eq. (4.15). Note that Eq. (5.8) entails solving Q optimization problems of the form in Eq. (5.6), however, the Q problems can be parallelized as they are completely independent. Decomposing the full μD spectrum reconstruction problem into Q subproblems effectively allows applying sparse recovery techniques, which in turn leads to a significant reduction of the number of measurements that are needed.

The value of Q is selected according to physical considerations and validated in practice, as described in Section 5.5. The μD vectors from Eq. (5.8) can be collected in sequences, one every δ slots, forming μD spectrograms of arbitrary length, depending on the specific application that is being performed, e.g., activity recognition, fall detection, gait segmentation, etc. In the following, we refer to the number of μD vectors considered in such spectrograms as Λ.

5.3.5 Sensing unit injection

SPARCS can exploit the sensing units in sparsely distributed communication packets to recover the μD spectrum of human movement. However, during communication between the AP and one or more terminals it may happen that the AP remains silent for longer than the duration of a processing window, WT_c, or that the received packets are fewer than the minimum number of measurements required for an accurate μD reconstruction. In these cases, the sparse recovery algorithm can not recover $\mathbf{H}_{\ell,b}(m)$ as the available sensing units are insufficient. To tackle this problem, we allow our system to *inject* sensing units into the channel whenever the number of communication packets is not sufficient for Alg. 5.1 to work. Different from existing ISAC frameworks, our sparse recovery approach allows us to introduce a minimal amount of overhead, as the μD spectrum can be recovered from a number of CIR samples which is much lower than the full length of the window W. Note that for the injection of a sensing unit it is sufficient to transmit the necessary CIR estimation fields, without any preamble and header as used in conventional packets, since the unit is only received at the AP itself and contains a known waveform.

Basis of the injection algorithm

In the following, we present the proposed injection procedure assuming that both communication packets and sensing units are transmitted at times that lie on a uniform grid with spacing T_c. This simplification is valid due to the fact that the slotted resampling process described in Section 5.3.1 is used. Therefore, we can describe the injection process in terms of windows of size W, where

Algorithm 5.2 Injection of sensing units in window m.

```
Input: $M_s$.
 1: # P1 - observation phase
 2: $N_a(m) \leftarrow$ no. of sensing units received in the first half of the window (either from reflected
    communications packets or injected).
 3: # P2 - scheduling phase
 4: $N_w(m) \leftarrow \max(M_s - N_a(m), 0)$.
 5: Schedule $\mathcal{S}_m = \{s_1, \ldots, s_{N_w(m)}\}$.
 6: # P3 - transmission phase
 7: for $q = mW/2, \ldots, (m+1)W/2 - 1$
 8:   if $q \in \mathcal{S}_m$
 9:     if no reflected comm. packet received
10:       Transmit the sensing unit.
11:       $\mathcal{S}_m \leftarrow \mathcal{S}_m \setminus \{q\}$.
12:     else
13:       Use the sensing unit from the comm. packet
14:       $\mathcal{S}_m \leftarrow \mathcal{S}_m \setminus \{q\}$.
15:     end if
16:   else
17:     if reflected comm. packet received
18:       Use the sensing unit from the comm. packet
19:       $\mathcal{S}_m \leftarrow \mathcal{S}_m \setminus \{\min_{s \in \mathcal{S}_m} s\}$.
20:     end if
21:   end if
22: end for
```

each value in the window occupies a *slot* which is a multiple of T_c. Due to slotted resampling, the slots can be empty if no packet was transmitted sufficiently close to it.

Our approach consists in setting a minimum number of sensing units *per window*, termed M_s, that allows a sufficiently accurate reconstruction of the μD signatures. We then transmit additional units whenever the number of reflections of communication packets in the window is not sufficient to meet this minimum requirement. The proposed method only requires the knowledge of whether a reflected communication packet is received in the *current* slot, i.e., no information about the future traffic pattern is needed.

Algorithm description

The algorithm, summarized in Alg. 5.2, operates in three phases, namely *observation* (P1), *scheduling* (P2) and *transmission* (P3). Recall that the μD extraction described in Section 5.3.2 follows a window-based approach, with subsequent windows overlapping by half of their length, as shown in Fig. 5.3. Consider a time instant between the end of window $m-1$ and the start of window $m+1$. This coincides with the *half* of window m, which is between slots $mW/2-1$ and $mW/2$. In this time instant we can observe how many reflected communication packets were received in the first half of window m, which spans the indices from $(m-1)W/2$ to $mW/2-1$ (P1, line 2 in Alg. 5.2). We denote this number as $N_a(m)$. The injection algorithm is executed

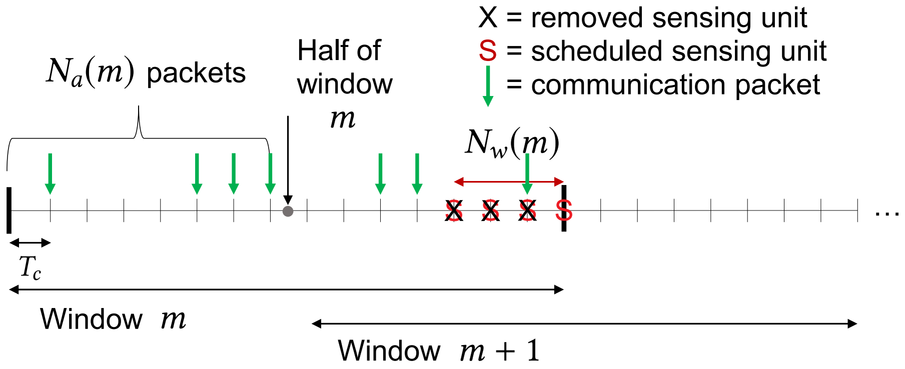


Figure 5.3: Example injection procedure with $M_s = 8, W = 16$. 4 sensing units are scheduled after P1 and P2. Then, as three reflected communication packets are received, we reuse them and the first three scheduled sensing units are not transmitted. The fourth sensing unit is instead injected in the last slot.

on a half-window basis at the time when window $m-1$ has ended and window $m+1$ has not yet started, as this allows reasoning on the sole current window m. Based on $N_a(m)$, we can compute how many sensing units we would need in the remaining half of window m in order to meet the requirement of at least M_s units, which we denote by $N_w(m) = \max(M_s - N_a(m), 0)$. However, the sensing process has no knowledge of when future communication packets will be received, so the best we can do is schedule the transmission of $N_w(m)$ sensing units in the next half-window. The slots in which these packets are scheduled can be selected according to a deterministic rule or a probability distribution. We call $\mathcal{S}_m = \{s_1, \dots, s_{N_w(m)}\}$ the set of indices of the slots in which we schedule the additional sensing units for the next half-window (P2, lines 4-5 in Alg. 5.2). While P1 and P2 are performed in a single time slot, before the second half of window m starts, P3 (lines 7-23 of Alg. 5.2) is a dynamic process that spans the whole second half of window m. The indices of the slots considered in this part of the algorithm are $q = mW/2, \dots, (m+1)W/2-1$. Note that some communication packets, of which we have no knowledge, may be received in this second half-window. The procedure iterates over the slots and in each of them checks if a sensing unit was scheduled for that slot, i.e., if $q \in \mathcal{S}_m$. There are four possible cases:

(**1**) $q \in \mathcal{S}_m$ and no communication packet was received in this slot. In this case we transmit the sensing unit, then remove q from $\mathcal{S}_m$.

(**2**) $q \in \mathcal{S}_m$ and a communication packet (or more) was received in this slot. In this case we reuse the sensing unit in the communication packet and remove q from $\mathcal{S}_m$.

(**3**) $q \notin \mathcal{S}_m$ and no communication packet was received in this slot. In this case we just move to the next slot without taking action.

(**4**) $q \notin \mathcal{S}_m$ and a communication packet (or more) was received in this slot. In this case we reuse the sensing unit in the communication packet, then we remove the next sensing unit from the scheduled ones, i.e., we set $\mathcal{S}_m \leftarrow \mathcal{S}_m \setminus \{\min_{s \in \mathcal{S}_m} s\}$.

Note that, despite operating on a half-window basis, due to the overlap of adjacent windows, our algorithm only poses a constraint on the minimum number of packets sent per *full window*. This means that half a window can be empty as long as enough sensing units are received in the other half.

Scheduling the sensing units

While the scheduling of the sensing units in P2 can be done with any arbitrary policy that guarantees that exactly $N_w(m)$ packets are scheduled in the next half window, we want to maximize the number of sensing units that can be piggybacked on communication packets, rather than using a dedicated transmission. From P3 in Alg. 5.2, one can see that scheduling the sensing units towards the end of the half-window leaves more time for possible communication packets to become available and thus be reused instead of injecting a new sensing unit. Consequently, in SPARCS we schedule the sensing units for the second half of window m as a burst of packets spaced by T_c, which occupy the last $N_w(m)$ slots of the window.

5.4 Implementation

In this section we describe the implementation of SPARCS on the mmWave SDR platform introduced in Section 4.5 of Chapter 4, basing our implementation on the IEEE 802.11ay Wi-Fi protocol, as it operates in the unlicensed 60 GHz band and supports CIR estimation for different BPs.

Testbed. As in RAPID, we use the FPGA-based baseband processor to generate, capture and process (custom or standard compliant) frames with up to 1.76 GHz of bandwidth. In the remainder of this chapter we use the same 60 GHz RF front-end used in RAPID. This simplifies experimentation as this is an unlicensed band, but we note that simply by changing the RF front-end, SPARCS can operate in a different band, e.g., for 5G-NR compatibility. The baseband processor supports various front-ends to operate in different frequency bands, e.g., at 28 GHz or 60 GHz [126]. To support the variable IFS extracted from real (or artificially generated) traces, we include a block RAM memory (BRAM) in the FPGA logic that stores the IFS that will be used in the experiments. The SM reads these values sequentially, introducing a delay in the system according to the value read from memory. The variable IFS functionality can be disabled at runtime to configure a *fixed* IFS. We remark that since we simultaneously use the up/down conversion stages from the *same* mmWave development kit, the Tx/Rx sub-systems are fed by the same local oscillator and thus the CFO is very low ($<$ 100 Hz), which enables the extraction of the μD values required by SPARCS.

IEEE 802.11ay CIR estimation details. In IEEE 802.11ay, *in-packet* beam tracking [25] is introduced, where the CIR is estimated using different BPs *within a single packet.* This is implemented by appending a given number of training (TRN) fields to the packet. A TRN field is composed of 6 TRN *units* formed by complementary Golay sequences of 128 BPSK modulated samples, for a total of 768 samples [111]. In our implementation, we use n_{TRN} TRN fields as the SPARCS sensing unit, where each TRN field employs a different BP, and n_{TRN} is the number of subjects being tracked by the system, as a single TRN field per subject suffices. Considering the typical number of people that are to be simultaneously tracked in human sensing systems,

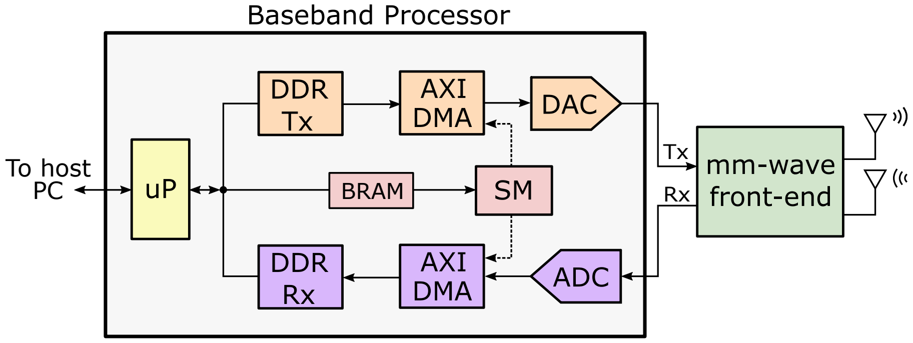


Figure 5.4: SPARCS implementation block diagram.

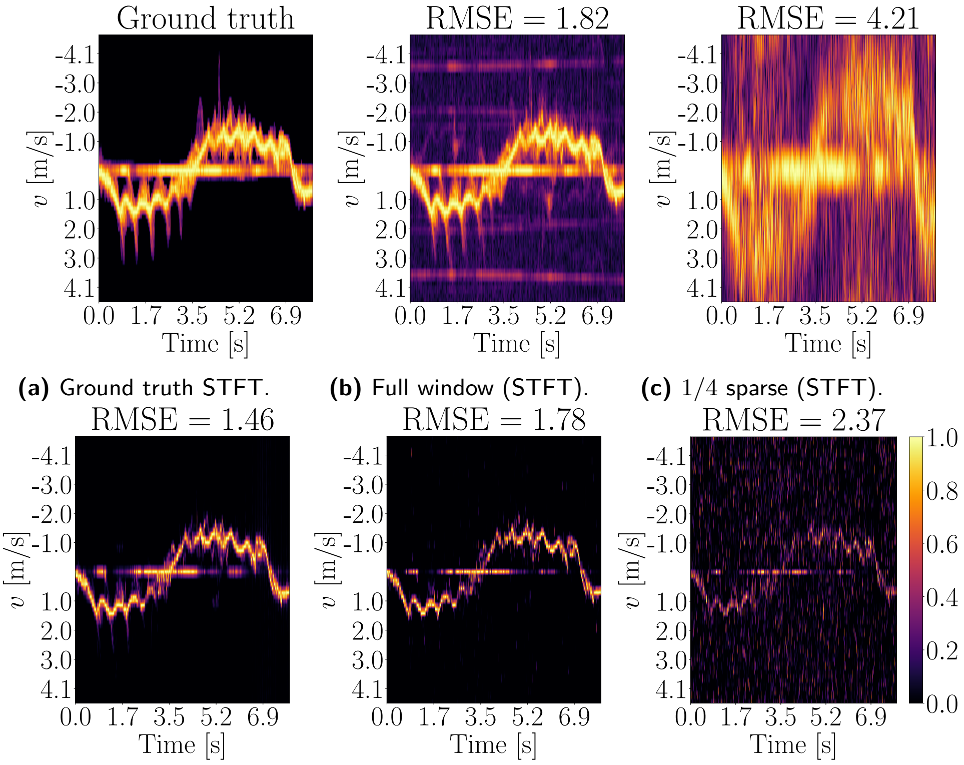
Table 5.1: Summary of the SPARCS implementation parameters. The suggested values based on experimental results are shown in bold.

System parameters		
Grid step	T_c	0.27 ms
Window length	W	64
Window shift	δ	32
Sparsity parameter	Ω	$\{1, 2, \mathbf{3}, 4, 5, 6, 7\}$
No. aggregated paths	Q	$\{1, 3, 5, 7, \mathbf{9}, 11, 13, 15\}$
Min. no. measurements	M_s	$\{4, \mathbf{8}, 16, 24, 32, 64\}$
IHT learning rate	η	1
IHT convergence threshold	ξ	10^{-4}
IHT maximum iteration number	$n_{\max}$	200

reasonable $n_{\rm TRN}$ values range from 1 to 10. The CIR estimates obtained from the TRN fields are then used as the input to SPARCS sparse recovery algorithm.

System parameters. In Tab. 5.1 we summarize the system parameters used in the implementation. We set $T_c = 0.27$ ms and $W = 64$, which lead to *(i)* a velocity resolution of $\Delta v = c/(2f_oWT_c) \approx 0.14$ m/s and *(ii)* aliasing-free velocity measurements up to $v_{\max} = \pm c/(4f_oT_c) \approx \pm 4.48$ m/s. These values are not critical to the functioning of our system, and can be modified according to specific implementation requirements. However, for reliable μD extraction without aliasing, it is advisable to adjust T_c to a value that allows capturing the range of velocities typically covered by human movement, e.g., approximately $\pm 2 - 3$ m/s for a walking person, and up to ±5 m/s for running or other fast movements [7]. Note that suitable values of T_c can also be obtained in 5G-NR systems, where a base station can transmit downlink CSI-RS frames with a periodicity between 0.3125 ms and 80 ms. For a 5G-NR carrier frequency of 28 GHz, using $T_c = 0.3125$ ms leads to $v_{\max} \approx \pm 8.57$ m/s, which is enough to capture fast human movement.

For people tracking, we use periodically transmitted in-packet beam training frames with 12 TRN units and antenna beams covering a FoV range from $-45°$ to $45°$. Then, we utilize the distance and AoA estimation procedure described in Section 4.3.3, as proposed in [31], to which



(a) Ground truth STFT. **(b)** Full window (STFT). **(c)** 1/4 sparse (STFT).

(d) Full window (SPARCS). **(e)** 1/4 sparse (SPARCS). **(f)** 1/16 sparse (SPARCS).

Figure 5.5: Walking μD spectrograms and RMSE for different levels of sparsity, obtained by uniformly removing samples for each window.

we refer for further details. We experimented with different values of M_s, Ω and Q, as reported in Tab. 5.1 and described in Section 5.5.3 and Section 5.5.5, while for the IHT algorithm we selected the parameters that led to the most accurate convergence results on our experiments, i.e., $\eta = 1$, $\xi = 10^{-4}$ and $n_{\max} = 200$.

5.5 Experimental results

We now present the experimental results obtained with our SPARCS testbed implementation. The experiments were performed in a laboratory of 6 × 7 meters with a complex multi-path environment due to additional reflections caused by furniture, computers, screens, and a wide whiteboard.

5.5.1 Results on synthetic traces

As a first qualitative result we show the μD spectrograms obtained by SPARCS on randomly sampled CIRs of a walking subject (see Fig. 5.5). For this, we use synthetic traces, generated by measuring the CIR using a uniform sampling interval equal to T_c, and then setting to 0 a

Table 5.2: Details of the 3 sequences of the pdx/vwave dataset.

Trace	Environment	No. frames	Duration
psu cs	University CS dept.	260326	1 : 00 h
library	Public library	1300671	4 : 00 h
ug	Coffee shop	895721	2 : 34 h

variable number of uniformly distributed values *per window* to simulate missing samples. This is a simplified case, as *(i)* the available (not removed) packets lie on a regular grid with spacing T_c, therefore no approximation error is introduced by slotted resampling, and *(ii)* samples are removed on a per-window basis, so a minimum number of packets in each window is guaranteed. Still, this evaluation is useful to highlight the impact of increasing the sparsity level of the measurements for SPARCS compared to standard STFT [5]. In the results presented in this section, no packet injection is performed, as we aim to assess the impact of the number of measurements per window on the reconstructed μD. In Fig. 5.5b we show the baseline walking spectrogram obtained using the standard STFT using the full window of 64 samples, as done in [31]. The spectrogram shows a typical walking μD modulation, with the contribution of the static clutter (the strong component at 0 velocity), of the torso (the strong oscillating component around ±1.5 m/s and the limbs (the faint contributions around the torso component). Moreover, a certain amount of noise and interference is present, as shown by the non-zero background level and the horizontal lines at around ±2 m/s and ±3.7 m/s. In Fig. 5.5c, the same method is applied to windows with only 16 out of 64 the samples retained, while the rest is set to 0. The impact is very strong as it completely corrupts the useful structure in the μD signature. From Fig. 5.5d to Fig. 5.5f we show the results obtained by SPARCS, on the same sequence, with $64, 16$ and 4 samples out of 64, respectively. At the top of each figure, we report the Root Mean-Squared Error (RMSE) of the μD with respect to a ground truth spectrogram, shown in Fig. 5.5a. This ground truth was obtained from the STFT output with full measurement windows (Fig. 5.5b), by manually isolating the useful μD spectrum containing the gait information and setting to 0 any background noise and interference lines. We observe two interesting aspects. On the one hand, the SPARCS algorithm can successfully recover the μD spectrum even when a large fraction of the samples is missing, and the quality of the result decreases *gracefully* with the sparsity of the available measurements. Unlike standard STFT, SPARCS almost completely eliminates the noise and interference in the estimated μD spectrogram. Such improvement is made possible by the sparsity constraint in Eq. (5.6), which allows for a lower RMSE than STFT operating on full measurement windows. This is the main reason why SPARCS not only reduces the overhead needed for human sensing, but also improves its accuracy.

5.5.2 Realistic traces: the pdx/vwave dataset

Next, we evaluate the performance of SPARCS on realistic Wi-Fi AP traces. This poses an experimental challenge, because commercial devices implementing the IEEE 802.11ay standard

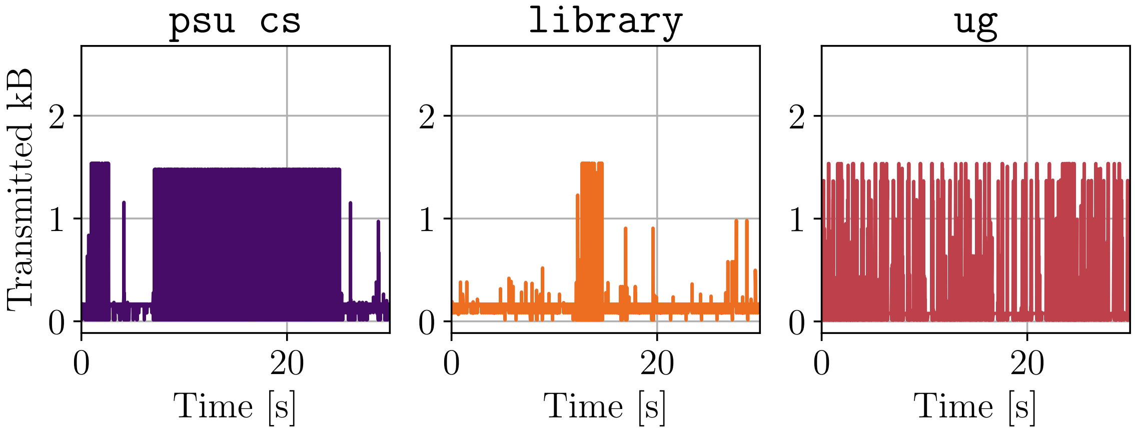


Figure 5.6: Example traffic patterns from the pdx/vwave dataset.

are not yet available, and no public datasets containing real traffic traces for the PHY layer of mmWave Wi-Fi (IEEE 802.11ay/ad) exist, to the best of our knowledge. For this reason, we used the pdx/vwave dataset, containing real traffic traces captured in different real environments from Wi-Fi APs employing a legacy (sub-6 GHz) Wi-Fi protocol [121]. Specifically, we use 3, over 1 hour long, traces from this dataset, called psu cs, library and ug, respectively. We select traces collected in different environments to represent different kinds of traffic patterns (see Tab. 5.2).

The pdx/vwave dataset includes information about the transmission instants and packet sizes of all packets outgoing from the considered AP. Exploiting this information, we perform our measurements transmitting packets according to these time patterns (see Section 5.5.3), using the BRAM in the FPGA to store the desired transmission instants (see Section 5.4). On top of the existing pdx/vwave communication patterns we use the injection algorithm (Alg. 5.2) to send additional sensing units when needed.

Even though the pdx/vwave dataset is based on a legacy sub-6 GHz Wi-Fi protocol, we argue that it is still reasonable to use it to obtain realistic packet transmission patterns. While in the pdx/vwave dataset the maximum physical layer PDU size is $\text{PPDU}_{\text{pdx}} = 1.5$ kB (without packet aggregation), in IEEE 802.11ay three main transmission modes are defined, namely High Throughput (HT), Directional Multi Gigabit (DMG) and Very High Throughput (VHT), with maximum physical layer PDU sizes, PPDU_{ay}, of 65 kB, 262 kB and 4692 kB, respectively [111], [117]. With the increase in the packet sizes, the data rates of mmWave systems have increased accordingly, and in IEEE 802.11ay they will range from 0.3 Gbps to several Gbps. As a numerical example, the traffic patterns in pdx/vwave with a typical bitrate of 4 Mbps would correspond to a bitrate of 0.7 Gbps in DMG IEEE 802.11ay when using an aggregated packet size of 262 kB instead of 1.5 kB. Note that traces with a larger number of packets and smaller PDU sizes (as will likely be the case in real deployments) will simply increase the sensing accuracy and further reduce the overhead.

5.5.3 Human activity recognition results

To evaluate the quality of the μD spectrograms extracted by SPARCS, we use them as the input to a HAR method. Specifically, we follow a standard approach, training a deep neural network on a dataset of $\Lambda \times W$ dimensional μD spectrograms, with $\Lambda = 200$ (equivalent to ≈ 1.76 s), in order to classify the movement performed by the person during that time. In order to provide a comparison with other IEEE 802.11ay HAR methods based on regular CIR sampling, such as our RAPID [31], presented in Chapter 4, we consider the 4 following activities: walking, running, sitting and waving hands. For HAR, we use a standard CNN architecture, composed of 4 *inception modules* [127] performing 1×1, 3×3 and 5×5 convolutions. The number of filters used is $8, 16, 32$ and 64 for the 4 modules, respectively. The convolutional blocks are followed by a fully-connected layer with 64 neurons, to which we apply Dropout [57], and a final Softmax layer with 4 outputs [53]. We use the *exponential-linear unit* activation function after each layer [54].

Training data

We collected a training dataset involving 6 different subjects performing the 4 activities, for a total duration of about 12 minutes each. This leads to over 400 partially overlapping, 1.76 s long, μD sequences per activity, which we then augmented as described shortly. Note that the training data only includes *uniformly sampled* CIR traces with sampling period $T_c = 0.27$ ms. The CNN training is done for 80 epochs, using a learning rate of 10^{-4}, the Adam optimizer and the cross-entropy loss function [53]. In order to enhance the robustness of the CNN, we apply an ad-hoc data augmentation strategy: we randomly remove some of the CIR samples in each window of the training dataset, and then apply SPARCS' IHT algorithm to reconstruct the spectrograms (see Section 5.5.1). We repeat the process using a sparsity level of $1/8, 1/4$ and $1/2$, enlarging the training dataset to 4 times its original size, for a total of approximately 1600 μD spectrograms per activity. A randomly selected subset of the training data (around 10 %) was used as a validation set to tune the CNN hyperparameters.

Test data

We test the CNN on the μD spectrograms obtained from CIR samples collected using the `pdx/vwave` packet traces described in Section 5.5.2. We collect four, randomly selected, 20 s long traces (one per activity) for each of the 3 sequences types (`psu cs`, `library` and `ug`). We repeat the experiments for different values of the minimum number of sensing units per window, $M_s = 4, 8, 16, 32, 64$, for a total of 60 test sequences. The test data involves a single subject, which was not included in the training set.

HAR F1 score

We evaluate the performance of the CNN with the *per-class* F1 score metric [128], which effectively summarizes the precision and recall and preserves the class-specific results. Fig. 5.7 shows the total average per-class F1 score over the 60 sequences, for different values of the minimum

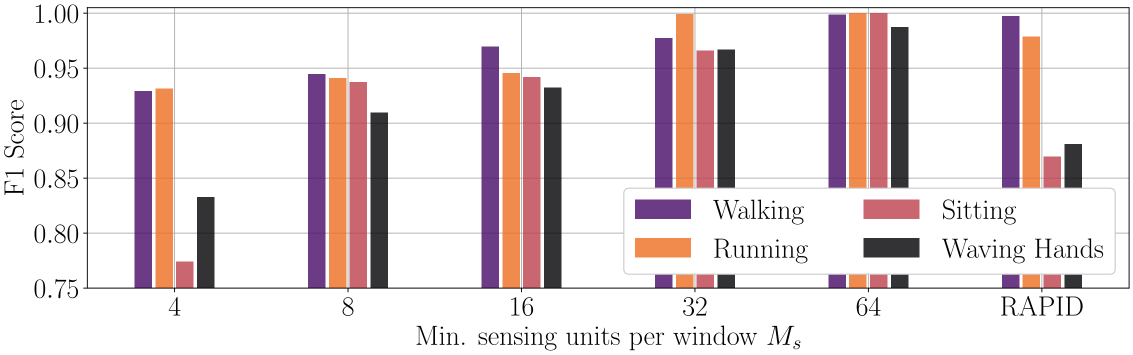


Figure 5.7: Per-class F1 scores obtained by SPARCS (for different values of M_s) and RAPID on our test dataset [31].

number of sensing units per window, M_s. As a baseline for comparison, we also report the F1 score obtained by the RAPID algorithm from [31], which extracts the μD signatures by regularly sampling the CIR. Our results show that SPARCS can reach over 0.9 F1 scores on all activities with $M_s = 8$ already, which corresponds to only 1/8 of the full measurements window. Notably, with $M_s = 4$, the low number of measurements per window affects significantly only the 'Sitting' and 'Waving hands' activities, which involve fine-grained movements and are therefore more difficult to classify. Finally, we compare the results from SPARCS and RAPID [31]. For a fair comparison, we implemented RAPID's STFT to extract the μD and trained the CNN on the resulting spectrograms without enlarging the dataset using different levels of sparsity described in the previous section. Instead, we directly use the training procedure of [31], since we found that the sparsity-based data augmentation slightly reduced RAPID's performance. It can be seen that SPARCS' sparse recovery problem formulation (Section 5.3.2) and enforcing a sparsity constraint on the individual paths is beneficial to HAR performance. The gap is particularly significant for 'Sitting' and 'Waving hands' as they involve lower energy traces in the spectrograms; these are more easily corrupted by noise and interference, that SPARCS is mostly able to reject (see, again, the comparison between Fig. 5.5b and Fig. 5.5d).

5.5.4 Overhead analysis

Increasing M_s to improve the HAR performance also increases the overhead of SPARCS. A first general measure of this can be obtained comparing the maximum size of a PPDU in IEEE 802.11ay to the size of a sensing unit. Recalling the three different modes introduced in Section 5.5.2 and the size of an IEEE 802.11ay TRN field (768 bits), we obtain that a sensing unit, with $n_{\text{TRN}} = 1$, is 0.1%, 0.03% and 0.002% of a PPDU in HT, DMG and VHT, respectively. Moreover, the channel occupation time for a sensing unit with $n_{\text{TRN}} = 1$ is 436 ns [111], which is a negligible fraction (0.16%) of a slot of duration T_c.

Next, to evaluate the overhead of SPARCS on a realistic communication scenario, we use the traces of the `pdx/vwave` dataset. In this way, we can also assess the impact of injecting sensing units, as they are not useful to the communication process. Denote by c_i the number of bits in

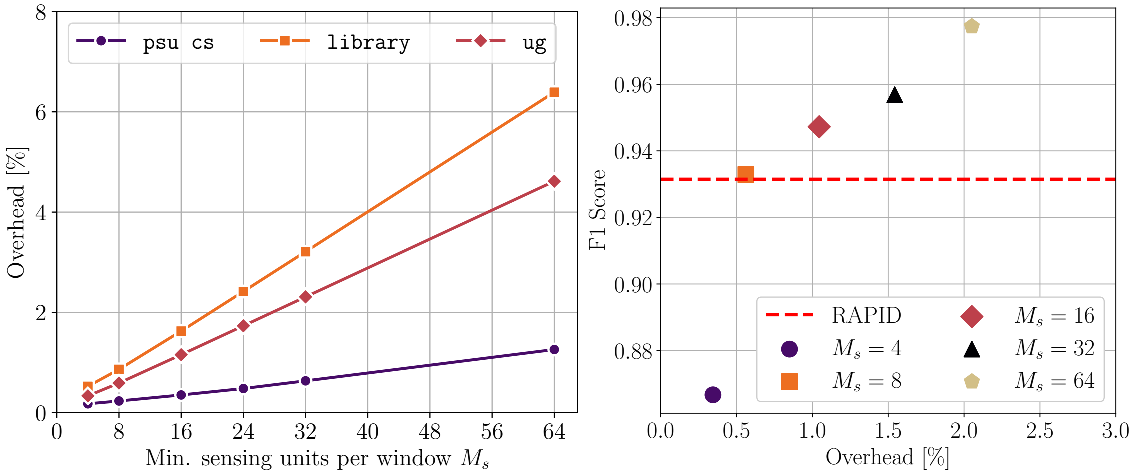


Figure 5.8: Overhead of SPARCS for different values of M_s in the three traces of the pdx/vwave dataset (left). Overhead vs. average HAR F1 score for different values of M_s (right).

the i-th communication packet transmitted in the trace. As the number of bits transmitted in each trace refers to a legacy, lower bitrate, Wi-Fi protocol, we rescaled the packet sizes according to the maximum PHY layer packet size in IEEE 802.11ay. We rescaled the size of packet i in each trace as $\tilde{c}_i = (\text{PPDU}_{\text{ay}}/\text{PPDU}_{\text{pdx}}) \times c_i$, with $\text{PPDU}_{\text{ay}} = 262$ kB. We call n_c the number of transmitted communication packets in a trace, by TRN_{len} the length, in bits, of a piggybacked or injected TRN field, n_{inj} the number of injected sensing units and n_{TRN} the number of TRN fields used in every sensing operation (we consider it fixed, whereas in reality it is determined by the number of subjects in the environment). We define the overhead as a function of n_{inj} as

$$\text{OH}(n_{\text{inj}}) = \frac{n_{\text{TRN}}\,(n_C + n_{\text{inj}})\,\text{TRN}_{\text{len}}}{\sum_{i=1}^{n_c} \tilde{c}_i}. \tag{5.9}$$

In Fig. 5.8, left, we show the overhead obtained on each of the three `pdx/vwave` traces, using $n_{\text{TRN}} = 1$. The overhead for different values of n_{TRN} can be obtained by using it as a multiplicative factor on the values in Fig. 5.8. We see that the overhead scales almost linearly as M_s is increased from 4 to 64. For values of $M_s < 32$, the entailed overhead is less than 4%, falling below 1% for $M_s = 8$ As a reference, we report the overhead for $M_s = 64$, which is the value obtained by injecting sensing units continuously into the channel, piggybacking them eventually on communication packets if possible. Note that existing approaches requiring uniform CIR sampling, like RAPID [31], would require an even higher overhead, as not only do they need 64 samples per window, but these samples have to be regularly spaced as no resampling procedure is carried out. This means they would have to take precedence over potential data packets so that they are sent exactly at the right sampling time.

From Fig. 5.8, right, one can see that SPARCS can achieve an F1 score of over 0.9 for every activity for a minimum of $M_s = 8$ sensing units per window, resulting in a sensing overhead of less than 1%. With this configuration, SPARCS achieves a better F1 score than existing approaches, while reducing overhead by a factor of 7 and being compatible with random access MAC protocols.

5.5.5 Sensitivity to the choice of the parameters

In Fig. 5.9, we show the effect of varying parameters Q, representing the number of paths aggregated around the person's position (see Eq. (5.8)), and Ω, which is the maximum number of resolvable Doppler components, equal to the sparsity parameter in the IHT algorithm. We computed the HAR per-class F1 score using a random subset of the 60 test sequences. The values adopted in our experiments are reported in Tab. 5.1.

Impact of changing Q

Our results show that SPARCS is robust to almost any value of Q when considering walking and running, whereas sitting and waving hands are negatively affected by reducing Q below 7. This is due to the fact that while walking and running are, in most cases, distinguishable even from the sole contribution of the torso, this is not true for sitting and waving hands that require including the reflection paths coming from the limbs. Computational complexity considerations are also in order for high values of Q, as it leads to solving Q times Eq. (5.6) at each μD extraction process. As the problems are independent, they can be solved in parallel, and thus a reasonable approach is to tune Q according to a trade-off between μD reconstruction accuracy and hardware resource availability for parallelization. In the following, we use $Q = 9$. Considering that we use $B = 1.76$ GHz transmission bandwidth (1 IEEE 802.11ay channel), the range resolution of SPARCS is $c/2B = 8.5$ cm. This means that summing the contribution of $\lfloor Q/2 \rfloor$ distance bins before and after the one corresponding to the torso, we include in the spectrum a region of ± 34 cm around the person's position, which is a reasonable value considering typical body sizes and that the subjects are moving.

Impact of changing Ω

Fixing $Q = 9$, in Fig. 5.9 (right), we show that the best values for Ω are 2 and 3 for all the activities. This is because using $\Omega = 1$ often leads to only reconstructing the 0 Doppler component in the spectrogram, losing the information on the person's movement. On the other hand, choosing Ω too high makes the IHT reconstruction imprecise, as with a low number of measurements per window enforcing more sparsity is beneficial to restrict the number of possible solutions to Eq. (5.6).

5.6 Related work

Dedicated mmWave radars. The high sensitivity of mmWaves to micro-Doppler shifts, together with DL methods for spectrogram analysis and classification, have been widely exploited to enable applications such as activity recognition [3], [4], person identification [5], [12] and bio-mechanical gait analysis [23]. The typical approach in these works is to transmit sequences of large bandwidth

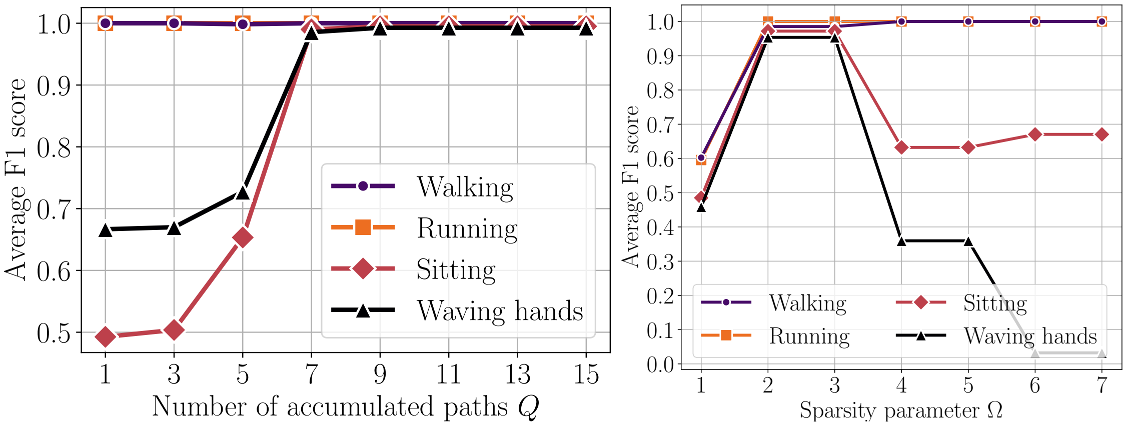


Figure 5.9: Per-class F1 scores aggregating a different number of paths Q (left) and changing Ω, the IHT sparsity parameter (right).

signals (of 2 to 4 GHz), with a rate dictated by the desired sensing resolution [11]. Thus, mmWave radar sensors have two main drawbacks:

(i) they are specifically tailored to sensing and cannot perform communication. Moreover, their deployment cost is relatively high as one single sensor can reliably cover a range of at most $8-10$ m due to the radial distortion and occlusion problems [90]. For this reason, ad-hoc radar sensor networks would need to be deployed in practical scenarios. Our method, in contrast, fully exploits existing mmWave communication systems with no modifications to the CIR estimation process or packet structures.

(ii) The fixed chirp transmission interval, which is related to the μD resolution, requires regular transmissions with continuous channel occupation. Some works have explored the possibility of randomly subsampling the chirp transmission intervals using compressive sensing [29] to either save computational resources [129] or reduce the effect of unwanted interference [130]. However, these works are based on a radar framework, where the transmission instants can be freely chosen and optimized. SPARCS, instead, reuses the given underlying communication traffic as much as possible and only injects small additional sensing units when necessary.

60 GHz Wi-Fi sensing. Research interest towards sensing with Wi-Fi devices working in the mmWave band has mostly focused on the 60 GHz IEEE 802.11ad/ay standards [28], [31], [96]. These works target various applications, such as person tracking and gesture recognition, exploiting CIR estimation to detect humans in the environment. However, they require dedicated and regular sensing signal transmissions in order to function properly, entailing a significant overhead and channel utilization for sensing.

In this chapter we significantly improve over the above-mentioned studies by enabling the reuse of randomly distributed communication packets via sparse recovery, whenever possible. This is of key importance to *integrate* sensing capabilities in communication devices while maintaining low overhead and complexity.

Integrated sensing and communication. A number of technical works address ISAC systems in next generation 5G/6G cellular networks [9], [27] and WLANs [10], [31]. Many of those target the joint communication and sensing waveform design [131] and are mostly oriented to automotive applications to measure distance and velocity of nearby vehicles. In contrast, few works focus on human sensing [27], which is the aim of the present chapter. All the above approaches alternate communication and sensing phases according to a time-division scheme, causing significant overhead and channel occupation. SPARCS instead, provides a full ISAC scheme, as it passively exploits communication traffic while dynamically injecting sensing units to cover silent periods. As a result, our method significantly reduces sensing overhead while at the same time improving the sensing accuracy.

5.7 Concluding remarks

In this chapter, we have designed and implemented SPARCS, the first mmWave ISAC system that can sense human μD signatures from irregular and sparse CIR estimates. These are obtained in a standard compliant way by both reusing optional CIR estimation fields appended to communication packets and sporadically injecting sensing packets whenever communication traffic is absent. Differently from the existing ISAC methods, SPARCS is based on a sparse recovery approach to the μD reconstruction, which is theoretically grounded in the intrinsic sparse multi-path environment of the mmWave channel. This enables an accurate μD extraction from a significantly lower number of randomly distributed CIR samples, thus drastically reducing the sensing overhead. After a CIR resampling step along the time domain, SPARCS performs an iterative sparse reconstruction in the frequency domain, decoupling different propagation paths at first, to leverage their sparsity property, and then combining them to obtain the final μD spectrum.

While SPARCS is compatible with different mmWave systems (e.g., 3GPP 5G-NR, and IEEE 60 GHz WLANs), for our implementation we used an IEEE 802.11ay SDR platform working in the 60 GHz band. We tested our system on a large set of standard-compliant CIR traces matching the traffic patterns of real Wi-Fi access points, performing a typical downstream application such as HAR. Our results show that SPARCS entails over 7 times lower overhead compared to prior methods, while achieving better performance.

6
Concluding remarks

Remote, unobtrusive sensing of human activity through RF signals holds the potential to become a key enabler for a variety of applications in healthcare, security, and monitoring of indoor and outdoor spaces. Two parallel approaches to the problem have been proposed by researchers. On the one hand, mmWave MIMO radar devices have been adopted to detect, track, and classify human activity in monitoring systems. The use of mmWaves provides fine-grained accuracy in capturing the features of human movement embedded in the reflected waveforms. However, radar sensors require ad-hoc installation and can only provide sensing functionalities, entailing high deployment costs. On the other hand, a large body of work has focused on Wi-Fi-based human activity sensing using commercial routers working in the sub-6 GHz band, to reuse existing communication hardware and avoid costly dedicated radar sensors. The downside of these systems is that they typically provide much lower accuracy than mmWave radars, due to their relatively low bandwidth and carrier frequency.

In this thesis, we made our contribution toward striking a balance between these two approaches. In Chapters 2 and 3, we focused on advanced mmWave radar sensing techniques to extract and recognize gait features from subjects walking in the environment. At first, we developed an integrated signal processing and deep learning method to recognize the identity of a person from their individual way of walking, proposing a solution that is both accurate, fast, and deployable on commercial edge computers. Secondly, we studied methods to improve the system, making it capable of *(i)* recognizing people on the fly with only a few seconds of observation time, and *(ii)* fusing the radar information with that obtained from a thermal camera, enabling joint body temperature screening, interpersonal distance monitoring, and contact tracing.

In the second part of the thesis, Chapters 4 and 5, we tackled the problem of Integrated Sensing And Communication (ISAC), exploring the use of mmWave communication networks as an appealing trade-off between mmWave radars and sub-6 GHz Wi-Fi sensing. In this sense, we presented an effective way to repurpose mmWave Wi-Fi, i.e., Wi-Gig, to perform radar-like sensing of human movement in indoor environments. The resulting system is capable of tracking multiple subjects at the same time while recognizing the activities they are performing, and their *identity* from gait characteristics. Then, we improved this system to fully integrate sensing with communication, proposing the first method to extract fine-grained micro-Doppler features of human movement from sparse and random communication traffic. For this, we leveraged the sparsity of the mmWave channel to reduce the sensing overhead and channel utilization, taking a step toward fully integrating sensing and communication.

In the following, we discuss future research directions stemming from the present work.

6.1 Future research directions

We identify three main research directions that require further development.

Networked sensing. The extension of the sensing paradigm to multiple, networked RF devices, can be an effective way to deal with the frequent blockage events that can happen at mmWave frequencies, especially when multiple subjects move in the same physical environment. This would allow covering bigger spaces, while also getting better results in the presence of occlusions. The networking aspect has to be explored for both ad-hoc radar devices and ISAC systems, which offer intrinsic communication capabilities. In this context, it is key to develop data fusion and collaborative sensing algorithms for sensors with overlapping fields of view, towards providing improved resilience to occlusions and better human tracking performance.

Multi-band ISAC. Future communication systems are expected to dynamically switch among different frequency bands to cope with deep mmWave channel fades and guarantee minimal communication service. A promising future direction is the design of novel strategies for ISAC that jointly leverage radio signals at different frequencies, i.e., on the sub-6 GHz and the mmWave bands of the radio spectrum. This allows combining sensing data with different characteristics, thus obtaining more detailed information on the surrounding environment that can be used to enhance system reliability and robustness. Specifically, high frequencies are very promising for sensing as they allow reaching high resolution by transmitting signals with wide bandwidth. However, such radio waves are blocked by obstacles in the propagation environment. In this case, signals at lower frequencies can help as they propagate through obstacles allowing one to still perform sensing, although at a lower resolution (due to the limited bandwidth).

Medium access control for ISAC. While there is a great research interest in PHY layer design for ISAC systems, little has been done yet to handle the competing needs of sensing and communication to access the *same* underlying radio channel. For this reason, we envision that an important challenge in the next few years will be the design of efficient and, possibly, distributed medium access control protocols that guarantee a fair and optimized sharing of the radio channel. In particular, the integration of beam management strategies that account for sensing metrics into communication protocols will be a key aspect to reach a full integration between sensing and communication.

References

[1] C.-H. Hsieh, Y.-F. Chiu, Y.-H. Shen, T.-S. Chu, and Y.-H. Huang, “A UWB radar signal processing platform for real-time human respiratory feature extraction based on four-segment linear waveform model,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 1, pp. 219–230, Feb. 2015.

[2] Y. Wang, H. Liu, K. Cui, A. Zhou, W. Li, and H. Ma, “m-Activity: Accurate and Real-Time Human Activity Recognition Via Millimeter Wave Radar,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Toronto, Ontario, Canada, Jun. 2021.

[3] G. Lai, X. Lou, and W. Ye, “Radar-Based Human Activity Recognition With 1-D Dense Attention Network,” *IEEE Geoscience and Remote Sensing Letters*, 2021.

[4] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, “Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar,” in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems (mmNets)*, Los Cabos, Mexico, Oct. 2019.

[5] J. Pegoraro, F. Meneghello, and M. Rossi, “Multiperson Continuous Tracking and Identification From mm-Wave Micro-Doppler Signatures,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021.

[6] J. Pegoraro and M. Rossi, “Real-Time People Tracking and Identification From Sparse mm-Wave Radar Point-Clouds,” *IEEE Access*, vol. 9, pp. 78 504–78 520, May 2021.

[7] B. Vandersmissen, N. Knudde, A. Jalalvand, *et al.*, “Indoor person identification using a low-power FMCW radar,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941–3952, 2018.

[8] P. Zhao, C. X. Lu, J. Wang, *et al.*, “mID: Tracking and Identifying People with Millimeter Wave Radar,” in *15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Santorini Island, Greece, May 2019.

[9] F. Liu, Y. Cui, C. Masouros, *et al.*, “Integrated sensing and communications: Towards dual-functional wireless networks for 6g and beyond,” *arXiv preprint arXiv:2108.07165*, 2021.

[10] F. Restuccia, *IEEE 802.11bf: Toward Ubiquitous Wi-Fi Sensing*, 2021. arXiv: 2103.14918.

[11] S. M. Patole, M. Torlak, D. Wang, and M. Ali, “Automotive radars: A review of signal processing techniques,” *IEEE Signal Processing Magazine*, vol. 34, no. 2, pp. 22–35, Mar. 2017.

[12] Z. Meng, S. Fu, J. Yan, *et al.*, “Gait Recognition for Co-Existing Multiple People Using Millimeter Wave Sensing,” in *AAAI Conference on Artificial Intelligence*, New York, New York, USA, Feb. 2020.

[13] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, “Micro-doppler effect in radar: Phenomenon, model, and simulation study,” *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.

[14] V. C. Chen, "Analysis of radar micro-doppler with time-frequency transform," in *Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing (Cat. No. 00TH8496)*, IEEE, 2000, pp. 463–466.

[15] A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–34, Apr. 2019.

[16] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Early Access 2021.

[17] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, California, USA, Jun. 2019.

[18] "Mobile Statistics Report, 2021-2025," Radicati Group Inc., 2021.

[19] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[20] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[21] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI based passive human activity recognition using attention based BLSTM," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2018.

[22] S. Mazahir, S. Ahmed, and M.-S. Alouini, "A survey on joint communication-radar systems," *Frontiers in Communications and Networks*, vol. 1, p. 9, 2021.

[23] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[24] J. A. Zhang, F. Liu, C. Masouros, *et al.*, "An Overview of Signal Processing Techniques for Joint Communication and Radar Sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021.

[25] Y. Ghasempour, C. R. C. M. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11ay: Next-Generation 60 GHz Communication for 100 Gb/s Wi-Fi," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 186–192, 2017.

[26] J. O. Lacruz, R. Ruiz, and J. Widmer, "A Real-Time Experimentation Platform for sub-6 GHz and Millimeter-Wave MIMO Systems," in *ACM MobiSys'21*, 2021.

[27] G. Li, S. Wang, J. Li, *et al.*, "Rethinking the Tradeoff in Integrated Sensing and Communication: Recognition Accuracy versus Communication Rate," *arXiv preprint arXiv:2107.09621*, 2021.

[28] F. Zhang, C. Wu, B. Wang, and K. R. Liu, "mmEye: Super-Resolution Millimeter Wave Imaging," *IEEE Internet of Things Journal*, vol. 8, no. 8, Apr. 2021.

[29] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications.* Cambridge University Press, 2012.

[30] M. Canil, J. Pegoraro, and M. Rossi, "milliTRACE-IR: Contact Tracing and Temperature Screening via mmWave and Infrared Sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 2, pp. 208–223, Apr. 2022.

[31] J. Pegoraro, J. O. Lacruz, E. Bashirov, M. Rossi, and J. Widmer, “RAPID: Retrofitting IEEE 802.11 ay Access Points for Indoor Human Detection and Sensing,” in *arXiv:2109.04819*, under submission, 2022.

[32] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, “SPARCS: A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems,” in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Milan, Italy, 2022.

[33] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, Jul. 2017.

[34] A. van den Oord, S. Dieleman, H. Zen, *et al.*, “WaveNet: A Generative Model for Raw Audio,” in *The 9th ISCA Speech Synthesis Workshop*, Sunnyvale, California, USA, Sep. 2016.

[35] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, “Radar-ID: human identification based on radar micro-Doppler signatures using deep convolutional neural networks,” *IET Radar, Sonar & Navigation*, vol. 12, no. 7, pp. 729–734, Jul. 2018.

[36] Y. Yang, C. Hou, Y. Lang, G. Yue, Y. He, and W. Xiang, “Person Identification Using Micro-Doppler Signatures of Human Motions and UWB Radar,” *IEEE Microwave and Wireless Components Letters*, vol. 29, no. 5, pp. 366–368, May 2019.

[37] S. Abdulatif, F. Aziz, K. Armanious, B. Kleiner, B. Yang, and U. Schneider, “Person identification and body mass index: A deep learning-based study on micro-Dopplers,” in *IEEE Radar Conference (RadarConf)*, Boston, Massachusetts USA, Apr. 2019.

[38] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, “Personnel recognition and gait classification based on multistatic micro-Doppler signatures using deep convolutional neural networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 669–673, May 2018.

[39] A. Jalalvand, B. Vandersmissen, W. De Neve, and E. Mannens, “Radar signal processing for human identification by means of reservoir computing networks,” in *IEEE Radar Conference (RadarConf)*, Boston, Massachusetts USA, Apr. 2019.

[40] V. Polfliet, N. Knudde, B. Vandersmissen, I. Couckuyt, and T. Dhaene, “Structured inference networks using high-dimensional sensors for surveillance purposes,” in *International Conference on Engineering Applications of Neural Networks (EANN)*, Crete, Greece, May 2018.

[41] J. Pegoraro, F. Meneghello, and M. Rossi, “Multi-Person Continuous Tracking and Identification from mm-Wave micro-Doppler Signatures,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021.

[42] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *ASME Transactions, Journal of Basic Engineering*, vol. 82, (Series D), no. 1, pp. 35–45, 1960.

[43] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin, *Principles of modern radar*. Raleigh, NC, USA: Scitech Publishing Inc., 2010.

[44] D. Lerro and Y. Bar-Shalom, “Tracking with debiased consistent converted measurements versus EKF,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 29, no. 3, pp. 1015–1022, Jul. 1993.

[45] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, Aug. 1996.

[46] T. Wagner, R. Feger, and A. Stelzer, "Radar signal processing for jointly estimating tracks and micro-Doppler signatures," *IEEE Access*, vol. 5, pp. 1220–1238, Feb. 2017.

[47] S. Bordonaro, P. Willett, and Y. Bar-Shalom, "Decorrelated unbiased converted measurement Kalman filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 2, pp. 1431–1444, Jul. 2014.

[48] G. Gennarelli, G. Vivone, P. Braca, F. Soldovieri, and M. G. Amin, "Multiple extended target tracking for through-wall radars," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6482–6494, Dec. 2015.

[49] J. W. Koch, "Bayesian approach to extended object and cluster tracking using random matrices," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 3, pp. 1042–1059, Oct. 2008.

[50] M. Feldmann, D. Franken, and W. Koch, "Tracking of extended objects and group targets using random matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1409–1420, Dec. 2010.

[51] Kuhn, Harold W, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[52] R. J. Fitzgerald, "Development of practical PDA logic for multitarget tracking by microprocessor," in *American Control Conference*, Seattle, Washington, USA, Jun. 1986.

[53] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* MIT press, 2016.

[54] D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.

[55] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015.

[56] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016.

[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[58] B. Jacob, S. Kligys, B. Chen, *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, Jun. 2018.

[59] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*, Graz, Austria, May 2006.

[60] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014.

[61] C. T. Nguyen, Y. M. Saputra, N. Van Huynh, *et al.*, "Enabling and emerging technologies for social distancing: A comprehensive survey," *arXiv preprint arXiv:2005.02816*, 2020.

[62] T. P. B. Thu, P. N. H. Ngoc, N. M. Hai, *et al.*, "Effect of the social distancing measures on the spread of COVID-19 in 10 highly infected countries," *Science of The Total Environment*, vol. 742, p. 140 430, 2020.

[63] F. de Laval, A. Grosset-Janin, F. Delon, *et al.*, "Lessons learned from the investigation of a COVID-19 cluster in Creil, France: effectiveness of targeting symptomatic cases and conducting contact tracing around them," *BMC Infectious Diseases*, vol. 21, no. 1, pp. 1–9, 2021.

[64] A. S. Ali and Z. F. Zaaba, "A study on contact tracing apps for covid-19: Privacy and security perspective," *Webology*, vol. 18, no. 1, 2021.

[65] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[66] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," *Computer Vision and Pattern Recognition*, 2018.

[67] M. I. Ribeiro, "Kalman and extended kalman filters: Concept, derivation and properties," *Institute for Systems and Robotics*, vol. 43, p. 46, 2004.

[68] S. Savazzi, V. Rampa, L. Costa, S. Kianoush, and D. Tolochenko, "Processing of body-induced thermal signatures for physical distancing and temperature screening," *IEEE Sensors Journal*, Early Access 2020.

[69] R. Faragher and R. Harle, "Location fingerprinting with bluetooth low energy beacons," *IEEE journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2418–2428, 2015.

[70] I. Galvan-Tejada, E. I. Sandoval, R. Brena, *et al.*, "Wifi bluetooth based combined positioning algorithm," *Procedia Engineering*, vol. 35, pp. 101–108, 2012.

[71] M. Cristani, A. Del Bue, V. Murino, F. Setti, and A. Vinciarelli, "The visual social distancing problem," *IEEE Access*, vol. 8, pp. 126 876–126 886, 2020.

[72] S. Bian, B. Zhou, H. Bello, and P. Lukowicz, "A wearable magnetic field based proximity sensing system for monitoring COVID-19 social distancing," in *Proceedings of the 2020 International Symposium on Wearable Computers*, 2020, pp. 22–26.

[73] M. Rezaei and M. Azarmi, "Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic," *Applied Sciences*, vol. 10, no. 21, p. 7514, 2020.

[74] A. J. Sathyamoorthy, U. Patel, Y. A. Savle, M. Paul, and D. Manocha, "Covid-robot: Monitoring social distancing constraints in crowded scenarios," *arXiv preprint arXiv:2008.06585*, 2020.

[75] N. Knudde, B. Vandersmissen, K. Parashar, *et al.*, "Indoor tracking of multiple persons with a 77 GHz MIMO FMCW radar," in *European Radar Conference (EURAD)*, Nuremberg, Germany, Oct. 2017.

[76] G. B. Dell'Isola, E. Cosentini, L. Canale, G. Ficco, and M. Dell'Isola, "Noncontact Body Temperature Measurement: Uncertainty Evaluation and Screening Decision Rule to Prevent the Spread of COVID-19," *Sensors*, vol. 21, no. 2, p. 346, 2021.

[77] C. Ferrari, L. Berlincioni, M. Bertini, and A. Del Bimbo, "Inner eye canthus localization for human body temperature screening," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 8833–8840.

[78] T. Lewicki and K. Liu, "AI thermometer for temperature screening: demo abstract," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 597–598.

[79] R. Zhang and S. Cao, "Extending reliability of mmwave radar tracking and detection via fusion with camera," *IEEE Access*, vol. 7, pp. 137 065–137 079, Sep. 2019.

[80] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, IEEE, 2019, pp. 1–7.

[81] M. Ulrich, T. Hess, S. Abdulatif, and B. Yang, "Person recognition based on micro-doppler and thermal infrared camera fusion for firefighting," in *21st International Conference on Information Fusion (FUSION)*, IEEE, 2018, pp. 919–926.

[82] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, "Learning longterm representations for person re-identification using radio signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, USA, Jun. 2020.

[83] Y. Cheng and Y. Liu, "Person reidentification based on automotive radar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, May 2022.

[84] "How to find the right Thermal imaging camera," *DIAS Infrared GmbH*, https://www.dias-infrared.de/pdf/How-to-find-the-right-thermal-imaging-camera_DIAS-Infrared.pdf, 2020.

[85] S. Prince, *Computer Vision: Models Learning and Inference.* Cambridge University Press, 2012.

[86] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, Feb. 2013.

[87] J. Nocedal and S. Wright, *Numerical optimization.* Springer Science & Business Media, 2006.

[88] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches.* John Wiley & Sons, 2006.

[89] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[90] L. Feng, S. Du, Z. Meng, A. Zhou, and H. Ma, "Evaluating mmWave Sensing Ability of Recognizing Multi-people Under Practical Scenarios," in *International Conference on Green, Pervasive, and Cloud Computing*, Springer, Dec. 2020, pp. 61–74.

[91] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, Springer, Amsterdam, Netherlands, Oct. 2016.

[92] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE conference on computer vision and pattern recognition (CVPR)*, Boston, Massachussetts, USA, 2015.

[93] H. T. Huynh and Y. Won, "Regularized online sequential learning algorithm for single-hidden layer feedforward neural networks," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1930–1935, Oct. 2011.

[94] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[95] P. Viola, M. Jones, *et al.*, "Robust real-time object detection," *International journal of computer vision*, vol. 4, no. 34-47, p. 4, Feb. 2001.

[96] C. Wu, F. Zhang, B. Wang, and K. R. Liu, "mmTrack: Passive multi-person localization using commodity millimeter wave radio," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, IEEE, 2020, pp. 2400–2409.

[97] H. Li, X. He, X. Chen, Y. Fang, and Q. Fang, "Wi-motion: A robust human activity recognition using WiFi signals," *IEEE Access*, vol. 7, pp. 153 287–153 299, 2019.

[98] F. Meneghello, D. Garlisi, N. D. Fabbro, I. Tinnirello, and M. Rossi, "Environment and Person Independent Activity Recognition with a Commodity IEEE 802.11 ac Access Point," *arXiv preprint arXiv:2103.09924*, 2021.

[99] 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)*, IEEE, 2017, pp. 1230–1239.

[100] B. Korany, H. Cai, and Y. Mostofi, "Multiple People Identification Through Walls Using Off-The-Shelf WiFi," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6963–6974, Apr. 2021.

[101] W. Jiang, C. Miao, F. Ma, *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 289–304.

[102] Z. Shi, J. A. Zhang, R. Y. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 540–554, Feb. 2022.

[103] A. Davoli, G. Guerzoni, and G. M. Vitetta, "Machine Learning and Deep Learning Techniques for Colocated MIMO Radars: A Tutorial Overview," *IEEE Access*, 2021.

[104] I. Pefkianakis and K.-H. Kim, "Accurate 3D Localization for 60 GHz Networks," New York, NY, USA: Association for Computing Machinery, 2018.

[105] S. D. Regani, C. Wu, B. Wang, M. Wu, and K. R. Liu, "mmWrite: Passive Handwriting Tracking Using a Single Millimeter Wave Radio," *IEEE Internet of Things Journal*, vol. 8, no. 17, Sep. 2021.

[106] Y. Ren, J. Lu, A. Beletchi, *et al.*, "Hand gesture recognition using 802.11 ad mmWave sensor in the mobile device," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, IEEE, 2021, pp. 1–6.

[107] F. Wang, F. Zhang, C. Wu, B. Wang, and K. R. Liu, "ViMo: Multi-person Vital Sign Monitoring using Commodity Millimeter Wave Radio," *IEEE Internet of Things Journal*, vol. 8, no. 3, Feb. 2021.

[108] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[109] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, Jun. 2016.

[110] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *11th international conference on information fusion (FUSION)*, IEEE, Cologne, Germany, Jun. 2008, pp. 1–6.

[111] IEEE 802.11 working group, "IEEE Draft Standard for Information Technology- Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks-Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications-Amendment: Enhanced Throughput for Operation in License-Exempt Bands Above 45 GHz," *IEEE P802.11ay/D3.0*, 2019.

[112] W.-C. Liu, F.-C. Yeh, T.-C. Wei, C.-D. Chan, and S.-J. Jou, "A Digital Golay-MPIC Time Domain Equalizer for SC/OFDM Dual-Modes at 60 GHz Band," en, *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 10, p. 10, 2013.

[113] D. Garcia, J. O. Lacruz, P. Jiménez Mateo, and J. Widmer, "POLAR: Passive object localization with IEEE 802.11ad using phased antenna arrays," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 1838–1847.

[114] D. Steinmetzer, D. Wegemer, M. Schulz, J. Widmer, and M. Hollick, "Compressive millimeter-wave sector selection in off-the-shelf IEEE 802.11 ad devices," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, 2017, pp. 414–425.

[115] J. O. Lacruz, D. Garcia, P. J. Mateo, J. Palacios, and J. Widmer, "mm-FLEX: An Open Platform for Millimeter-Wave Mobile Full-Bandwidth Experimentation," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '20, Toronto, Ontario, Canada: Association for Computing Machinery, 2020, pp. 1–13.

[116] SIVERSIMA, *EVK06002 Development Kit*, `https://www.siversima.com/product/evk-06002-00/`, 2020.

[117] IEEE 802.11 working group, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band.," *IEEE Standard 802.11ad*, 2012.

[118] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free Your CSI: A Channel State Information Extraction Platform For Modern Wi-Fi Chipsets," in *Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH)*, Los Cabos, Mexico, Oct. 2019.

[119] Y. Zheng, Y. Zhang, K. Qian, *et al.*, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Seoul, Republic of Korea, 2019.

[120] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *ICLR: International Conference on Learning Representations*, 2015, pp. 1–15.

[121] C. Phillips and S. Singh, *CRAWDAD dataset pdx/vwave (v. 2009-07-04)*, Available at `https://crawdad.org/pdx/vwave/20090704/wlan_nano_fcs`, Jul. 2009.

[122] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3gpp nr at mmwave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[123] P. Kumari, J. Choi, N. González-Prelcic, and R. W. Heath, "Ieee 802.11 ad-based radar: An approach to joint vehicular communication-radar system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3012–3027, 2017.

[124] P. Babu and P. Stoica, "Spectral analysis of nonuniformly sampled data–a review," *Digital Signal Processing*, vol. 20, no. 2, pp. 359–378, Mar. 2010.

[125] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[126] Sivers Semiconductors, *Evaluation Kits*, `https://www.sivers-semiconductors.com/sivers-wireless/evaluation-kits`, 2021.

[127] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachussets, USA, Jun. 2015.

[128] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th Conference on Message Understanding*, ser. MUC4 '92, McLean, Virginia: Association for Computational Linguistics, 1992, pp. 22–29.

[129] S. Stanković, I. Orović, T. Pejaković, and M. Orović, "Compressive sensing reconstruction of signals with sinusoidal phase modulation: Application to radar micro-doppler," in *2014 22nd Telecommunications Forum Telfor (TELFOR)*, IEEE, 2014, pp. 565–568.

[130] E. Sejdić, I. Orović, and S. Stanković, "Compressive sensing meets time–frequency: An overview of recent advances in time–frequency processing of sparse signals," *Digital signal processing*, vol. 77, pp. 22–35, 2018.

[131] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, "Toward dual-functional radar-communication systems: Optimal waveform design," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4264–4279, 2018.

List of publications

Journals

[5] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson Continuous Tracking and Identification From mm-Wave Micro-Doppler Signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021.

[6] J. Pegoraro and M. Rossi, "Real-Time People Tracking and Identification From Sparse mm-Wave Radar Point-Clouds," *IEEE Access*, vol. 9, pp. 78 504–78 520, May 2021.

[30] M. Canil, J. Pegoraro, and M. Rossi, "milliTRACE-IR: Contact Tracing and Temperature Screening via mmWave and Infrared Sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 2, pp. 208–223, Apr. 2022.

[31] J. Pegoraro, J. O. Lacruz, E. Bashirov, M. Rossi, and J. Widmer, "RAPID: Retrofitting IEEE 802.11 ay Access Points for Indoor Human Detection and Sensing," in *arXiv:2109.04819*, under submission, 2022.

[132] J. Pegoraro, M. Canil, A. Shastri, P. Casari, and M. Rossi, "ORACLE: Occlusion-Resilient And self-Calibrating mmWave Radar Network for People Tracking," in *arXiv:2208.14199*, under submission, 2022.

Conferences

[32] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "SPARCS: A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems," in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Milan, Italy, 2022.

[133] J. Pegoraro, D. Solimini, F. Matteo, E. Bashirov, F. Meneghello, and M. Rossi, "Deep Learning for Accurate Indoor Human Tracking with a mm-Wave Radar," in *2020 IEEE Radar Conference (RadarConf20)*, Florence, Italy, Sep. 2020.

[134] A. Shastri, M. Canil, J. Pegoraro, P. O. Casari, and M. Rossi, "mmSCALE: Self-Calibration of mmWave Radar Networks from Human Movement Trajectories," in *2022 IEEE Radar Conference (RadarConf22)*, New York, USA, Mar. 2022.

[135] J. Pegoraro and M. Rossi, "Human Tracking with mmWave Radars: a Deep Learning Approach with Uncertainty Estimation," in *2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*, Oulu, Finland, 2022.

Patent Applications

[136] A. Shastri, M. Canil, J. Pegoraro, P. O. Casari, and M. Rossi, "Method for self-calibration of mmWave radar networks," *Italian Patent Application*, no. 812022000069836, Mar. 2022.

[137] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "Method and system for the sparse reconstruction of the micro-Doppler spectrum in joint communication and sensing applications," *Italian Patent Application*, no. 102022000008906, May 2022.