Hand Gesture Recognition for an Off-the-Shelf Radar by Electromagnetic Modeling and Inversion

ARTHUR SLUŸTERS, Université catholique de Louvain, LouRIM, Belgium
SÉBASTIEN LAMBOT, Université catholique de Louvain, Earth and Life Institute, Belgium
JEAN VANDERDONCKT, Université catholique de Louvain, LouRIM, Belgium

Microwave radar sensors in human-computer interactions have several advantages compared to wearable and image-based sensors, such as privacy preservation, high reliability regardless of the ambient and lighting conditions, and larger field of view. However, the raw signals produced by such radars are high-dimension and relatively complex to interpret. Advanced data processing, including machine learning techniques, is therefore necessary for gesture recognition. While these approaches can reach high gesture recognition accuracy, using artificial neural networks requires a significant amount of gesture templates for training and calibration is radar-specific. To address these challenges, we present a novel data processing pipeline for hand gesture recognition that combines advanced full-wave electromagnetic modelling (EM) and inversion with machine learning. In particular, the physical model accounts for the radar source, radar antennas, radar-target interactions and target itself, *i.e.*, the hand in our case. To make EM processing feasible, the hand is emulated by an equivalent infinite planar reflector, for which analytical Green's functions exist. The hand, located at a specific distance from the radar, is therefore characterized by an apparent dielectric permittivity. This apparent permittivity depends on the hand only (e.g., size, electric properties, orientation) and, together with the distance, determines wave reflection amplitude. Through full-wave inversion of the radar data, the physical distance as well as this apparent permittivity are retrieved, thereby reducing by several orders of magnitude the dimension of the radar dataset, while keeping the essential information. Finally, the estimated distance and apparent permittivity as a function of gesture time are used to train the machine learning algorithm for gesture recognition. This physically-based dimension reduction enables the use of simple gesture recognition algorithms, such as template-matching recognizers, that can be trained in real time and provide competitive accuracy with only a few samples. We evaluate significant stages of our pipeline on a dataset of 16 gesture classes, with 5 templates per class, recorded with the Walabot, a lightweight, off-the-shelf array radar. We also compare these results with an ultra wideband radar made of a single horn antenna and lightweight vector network analyzer, and a Leap Motion Controller.

CCS Concepts: • **Human-centered computing** → **Gestural input**; **Graphical user interfaces**; *Interactive systems and tools*; • **Computing methodologies** → *Model development and analysis*; • **Software and its engineering** → Runtime environments; • **Hardware** → Radio frequency and wireless interconnect.

Additional Key Words and Phrases: Dimension reduction, Gesture-based interfaces, Hand gesture recognition, Mid-air gestural interaction, New datasets, Radar-based interaction.

ACM Reference Format:
Arthur Sluÿters, Sébastien Lambot, and Jean Vanderdonckt. 2022. Hand Gesture Recognition for an Off-the-Shelf Radar by Electromagnetic Modeling and Inversion. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3490099.3511107

1 INTRODUCTION

3D gesture-based User Interfaces (UIs) [32], which promise a natural and intuitive interaction [25], typically rely on motion inputs [39] recognized by computer vision or sensor inputs captured by wearable devices [24]. More specifically for hand gesture recognition (see [15] for a survey), computer vision acquires raw data from non-wearable, image-based devices like Ultraleap Leap Motion Controller (LMC), Intel RealSense, Microsoft Kinect Azure, or Duo3D, to accurately model and recognize hand gestures [18]. Consequently, they may suffer from problems inherent to image-based processing [8, 9, 22, 52, 54]: sensitivity to ambient conditions, particularly lighting, limited field of view, transient or permanent vision occlusion, and privacy concerns raised by a visible device observing the end user. On the other hand, wearable devices, such as smart rings [20], smartwatches, armbands, or mobile devices featuring Inertial Measuring Units (IMUs, *i.e.*, groups of sensors, such as accelerometers, gyroscopes, and magnetometers [3, 37]) are able to stream raw data in real-time to track gestures. Although these devices offer high precision in data streaming, they do not all include self-calibration (*e.g.,* the Thalmic Myo armband requires muscular calibration on the end user's forearm) and are considered obstrusive [8], artificial, less ecologically valid, and more intrusive.

Radar sensing technologies [52] recently appeared as an alternative to these techniques as they do not suffer from the aforementioned limitations while maintaining acceptable accuracy. They have been successfully applied in various domains such as virtual reality [26], activity recognition [8, 9], material recognition [19, 52], and tangible interaction [53]. Prior works on gesture recognition using radar sensing, such as [7, 9, 22, 40], typically rely on a fixed, custom radar which is hard to reproduce, and sophisticated Machine Learning (ML) or Deep Learning (DL) algorithms to cope with the high dimensionality of radar signals. These works have not yet been transposed to mobile, off-the-shelf radars. For instance, the Walabot device, while being commercially available and deployable in various mobile and stationary contexts of use, induces a new series of constraints imposed by its limited amount of antennas, its small size, and its relatively narrow bandwidth, thus making hand gesture recognition with this device an unsolved problem. An exception is perhaps the Google Soli chip [36], which has been integrated into a smartphone for recognizing various classes of gestures: it is therefore mobile and available. In any case, the high dimensionality of radar signals, sometimes with sparsity [34], remains an open problem that could significantly impact gesture recognition [46].

To address the aforementioned limitations and challenges, this paper presents a novel software pipeline composed of eight stages for hand gesture recognition with the goal of *dimension reduction* [45] by full-wave electromagnetic modeling and inversion: the high-dimensional format of radar signals is transformed into a bi-dimensional space made up of only two physically-meaningful features: the distance between the end-user and the radar and the *apparent permittivity* [14], denoted by ϵ, a quantity that determines polarizability of a dielectric material (for example, a material with high permittivity polarizes more under an electric field than a material with low permittivity). This dimension reduction allows the use of simpler yet more efficient gesture recognizers, such as template-matching [13] gesture recognizers, such as Jacknife [44], which support adding new gestures in real-time with a few training templates, thus enabling users to create their own gesture sets and introduce user-defined gesture sets [21].

The remainder of this paper is organized as follows. Section 2 reviews briefly existing works on hand gesture recognition and more extensively on radar sensing for hand gesture interaction in Human-Computer Interaction (HCI). It also describes the specifications of the Walabot, the mobile off-the-shelf radar used in this paper, and justifies its usage. Section 3 then describes the first contribution: our novel software pipeline for hand gesture recognition for dimension reduction into distance and apparent permittivity by applying two major techniques: full-wave electromagnetic modeling and inversion. Section 4 explains and justifies the second contribution: our dataset with 16 gesture classes and only 5

templates per class, as well as its recording procedure executed simultaneously with the Walabot (our target device), a horn (a custom-built radar with a single antenna as a comparison point in radar sensing), and an LMC (as a comparison point for hand gesture recognition outside the area of radar sensing). Section 5 evaluates the significant stages of our pipeline on the acquired dataset in terms of recognition rate, execution time, and vector size of input signals, and compares these results with those obtained for the horn and the LMC. Section 6 discusses the results of our evaluation and their implications for the design of hand gesture recognition systems based on off-the-shelf radars. These insights form our third contribution. Section 7 reflects on the limitations of this work and discusses how to address them. Finally, Section 8 concludes this paper and discusses future avenues for research in radar gesture recognition.

2 RELATED WORK

This section reviews existing work related to hand gesture recognition, first in general with or without any wearable device (Section 2.1), then with a focus on radar sensing in HCI (Section 2.2). Finally, Section 2.3 describes the specifications of the Walabot device, the mobile off-the-shelf radar used in this paper, and justifies its usage.

2.1 Hand Gesture Recognition

For several years, many vision-based sensors have become available at affordable prices, such as Intel RealSense camera, Microsoft Kinect Azure, or UltraLeap Leap Motion Controller (LMC). The latter two have been particularly popular in the literature, with many applications deployed in research and development. The LMC, a cheap off-the-shelf sensor (~100$) that can be plugged via USB into a computer, is now recognized as a robust and precise device for hand tracking [49], particularly with Deterministic Learning [55]. It is suitable for 3D mid-air gesture interaction (see [15] for a survey on 3D gesture interaction, [28] for a survey in mid-air interaction, and [51] for a systematic literature review on hand gesture recognition) as it tracks two hands and their ten fingers [17], including the palm vector and hand radius. This sensor is comprised of two cameras and three infrared (IR) LEDs. The interaction space is 0.6 m (1.97 ft) wide by 0.6 m long by 0.6 m deep, resulting in an 0.22 m^3 (7.76 ft^3) volumetric space in the shape of an inverted pyramid above the sensor. The LMC, as well as other devices belonging to the same family, is subject to several factors: sensitivity to ambient conditions, particularly lighting [54] (*e.g.,* the LMC should not be operated in a bright or reflective environment), limited field of view [22] (*e.g.,* a clear view should be maintained between the LMC and the end-user), vision perturbation [12] (*e.g.,* any occlusion, even temporary, may affect signal processing and imply some recalibration), and privacy concerns [8] (*e.g.,* a device visible to the end-user may raise concerns related to trust, security, privacy, and social acceptance as the end-user fears to experience some trouble with the device in front of others).

Other types of sensors that aim at solving (some of) these issues are being developed, such as Li *et al.*'s Aili [35], a custom device that provides hand skeleton data without the privacy concerns of vision-based sensors. The device acts as a table lamp and features 288 LEDs hidden in its lampshade. The 16 photodiodes placed in its base can determine which LEDs are obstructed by the hand. This information is then used to reconstruct a hand skeleton with reasonable accuracy. In addition, sensors such as utilizing Electromyography (EMG) or Inertial Measurement Units (IMUs) have also been used for gesture recognition [24, 37]. Although wearable devices should be worn and, as such, always induce some intrusive character that could be accepted or rejected, robust algorithms exist for gesture segmentation, recognition, and interaction [38, 44]. For example, Akl *et al.* [3] relied on Dynamic Time Warping (DTW) [39] and affinity properties to recognize gestures based on accelerometers. Laput and Harisson [31] used accelerometer and bio-acoustic data from an off-the-shelf smartwatch to recognize a set of 25 hand activities. Desmedt *et al.* [18] used a depth and skeletal model in their gesture dataset. Li *et al.* [33] recognized finger gestures with high precision using WE-kNN algorithm. 3D

gesture recognition have been made invariant to space and rotation [5] to become insensitive to variations of gesture articulations performed in space. Such 3D gestures have been also incorporated in some design tools, such as Magic V1.0 [6] and Magic V2.0 [27], and development environments, such as CodeSpace [11].

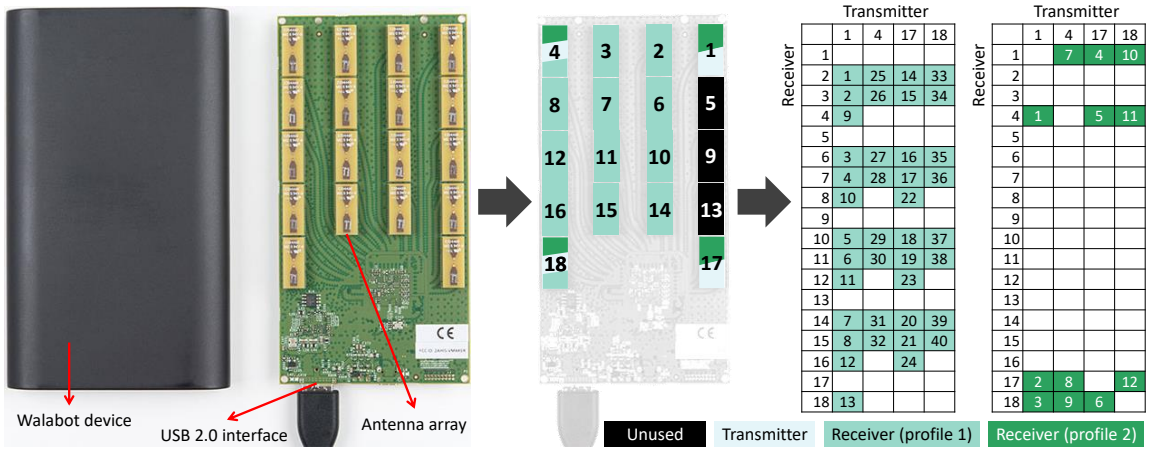
2.2 Radar-based Interaction

In principle, radar sensing techniques allow interaction without any wearable and visible device since a radar can be operated below a surface such as a desk [8], behind a wall, and behind different materials without significantly affecting the recognition [40] (*e.g.,* wallpaper, cardboard, and wood benefit from relatively low permittivity). Radars are also insensitive to weather and lighting conditions [54].

The history of radar-based interaction starts with the Magic Carpet [41], a Doppler radar used for sensing coarse 3D body motions to be recognized by signal processing techniques. Since then, radar-based interaction has become an increasingly popular research topic. For example, RadarCat [52] recognizes physical objects and material placed on top of the sensor in real-time by extracting signals from a radar and classifying them using a random forest classifier. In their study, 26 materials, 16 transparent materials, and 10 body parts from 6 participants were accurately recognized. This shows the real potential of permittivity, since different materials used in this study, with various physical properties such as thickness, were properly recognized. Transparent materials offer in general an excellent, i.e., low, permittivity that does not alter recognition. Yeo *et al.* [53] used a radar in the context of tangible interaction for counting, ordering, and identifying objects involved in a tangible setup, for tracking their orientation, movement, and between-object distance, three variables that were originally captured by infra-red [43]. Beyond object and material detection and classification, radars are starting to be widely used in several domains of application, such as indoor human sensing with commodity radar [4], human activity recognition [8], human position estimation [9], motion detection and classification [40].

GestureVLAD [10] is a framework for gesture recognition with Doppler radar which supports slight variations in gesture execution while accurately differentiating between gesture classes and is fast enough for real-time recognition. Pantomime [40] mount a fixed feet-based radar with a high-frequency *i.e.*, 76-81 GHz equipped with a 4 GHz continuous bandwidth and composed of 4 receiving antennas and 3 transmitting antennas. Deep learning, *i.e.*, LSTM and Pointnet++, is used to accurately recognize 21 gestures acquired by 45 participants from 3D point clouds obtained from the radar. This radar works at a frequency that is about ten times larger than the Walabot device, thus resulting in a wavelength ten times smaller and a resolution about ten times finer than the Walabot. Although their radar is limited in the number of antennas (7), they can be oriented towards a better lateral resolution, which is not the case with rigid radars, such as the Walabot. Similarly, Wang *et al.* [47] require only two antennas in their radar to recognize 2D stroke gestures: their low-dimensionality, as opposed to 3D gestures, do not require more antennas. Short-range radar-based gestures could also be recognized using 3D convolutional neural networks (CNNs) with a triplet loss [23].

Most existing works exploit a custom radar built with specific, on-purpose features, a system that is hard to reproduce and to reuse unless one benefits from extensive experience in electronics, radar sensing, and antenna management, thus presenting a barrier to direct adoption in mainstream radar-based interaction. In contrast, the Google Soli [36, 48], one of the few off-the-shelf available radar sensors specifically designed for interactive applications, initially tracks micro gestures, such as finger wiggle, hand tilt, check mark, or thumb slide. It has then been expanded to gesture recognition by combining deep convolutional and Recurrent Neural Networks (RNN) on a dataset of 11 gesture classes with a high-frequency (60 GHz), short-range radar-based sensing. More recently, Choi *et al.* [16] proposed a gesture recognition system for the Google Soli able to detect and recognize a set of 10 hand gestures in real-time. In principle, the Google Soli could be embedded in any mobile device, such as a smartwatch or a smartphone, or in any physical



Receiver \ Transmitter	1	4	17	18
1				
2	1	25	14	33
3	2	26	15	34
4	9			
5				
6	3	27	16	35
7	4	28	17	36
8	10		22	
9				
10	5	29	18	37
11	6	30	19	38
12	11		23	
13				
14	7	31	20	39
15	8	32	21	40
16	12		24	
17				
18	13			

Receiver \ Transmitter	1	4	17	18
1		7	4	10
2				
3				
4	1		5	11
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17	2	8		12
18	3	9	6	

Fig. 1. The Walabot Developer: device, antenna array, antenna IDs, and antenna pairs IDs used in profiles 1 and 2.

object. For instance, Flintoff *et al.* [19] integrated Soli into a two-fingered robot hand for object classification. Hayashi *et al.* developed RadarNet, an algorithm optimized for efficiently recognizing five gesture classes (*i.e.*, left/right/up/down swipe and an omnidirectional swipe) on computationally constrained battery-powered devices. Attygale *et al.* [7] also exploited a Google Soli sensor via a three-dimensional convolutional neural network (Conv3D) and a spectrogram-based ConvNet to recognize on-object gestures (*e.g.*, 94% for a five-gesture set). Unlike embedding a radar in any object [53], an external radar enables any physical object to be tracked. Finally, Ren *et al.* [42] explored the possibility of combining radar gesture sensing and wireless communication on a smartphone equipped with an 802.11ad 60GHz WiFi chipset, thus avoiding the need for a dedicated radar sensor for gesture recognition. Despite the sub-optimal positioning of the WiFi antennas for gesture sensing, their system achieved high accuracy on a dataset of five gestures.

Radar-based datasets, such as [2, 40], start to be built but remain rarely publicly accessible. When they are, their size is in the order of several GigaBytes, which makes them challenging to process.

2.3 Walabot-based Interaction

The Walabot Developer consists of a compact off-the-shelf ultra-wideband (UWB) frequency-modulated continuous-wave (FMCW) radar. Its dimensions are 144 mm×85 mm×18 mm (5.67 in.×3.35 in.×0.71 in.). This device could be coupled to a smartphone, a tablet, or a computer via a single USB cable, which makes it appropriate for both (semi-)mobile and stationary contexts of use. It exists in two versions: a US/FCC version operating in the 3.3-10 GHz frequency range and an EU/CE version operating over the narrower 6.3-8 GHz range. The latter is thus the most restrictive and challenging for gesture recognition. We will therefore use this version in the rest of this paper. The Walabot Developer consists of an array of 18 antennas, including four used as transmitters (depicted in orange in Fig. 1), the rest being only receivers (depicted in green in Fig. 1). Depending on the configuration, it senses motion data with up to 40 pairs of antennas. The Walabot SDK provides two different profiles for distant scanning, which define the set of antenna pairs (right part of Fig. 1), the number of fast-time samples per frame, and the frame rate:

- Profile 1 (PROF_SENSOR): 40 antenna pairs, 8192 fast-time samples/frame, approximately 20 frames/s
- Profile 2 (PROF_SENSOR_NARROW): 12 antenna pairs, 4096 fast-time samples/frame, approximately 41 frames/s

We selected the Walabot as it has been proved effective and efficient in various domains of applications, including material identification [1] and activity recognition [8, 58]. For example, Avrahami *et al.* [8] are able to recognize human

Recognition	(a)		(b)		(c)		(d)	
	Walabot	Horn	Walabot	Horn	Walabot	Horn	Walabot	Horn
Recognition rate [%]	79.7	94.6	84.5	95.4	62.4	55.9	70.7	54.5
Time [ms]	7.3	3.8	5.8	3.3	0.4	0.2	0.5	0.2
Gesture size [kB]	212.5	111.25	170	111.25	5	1.25	5	1.25

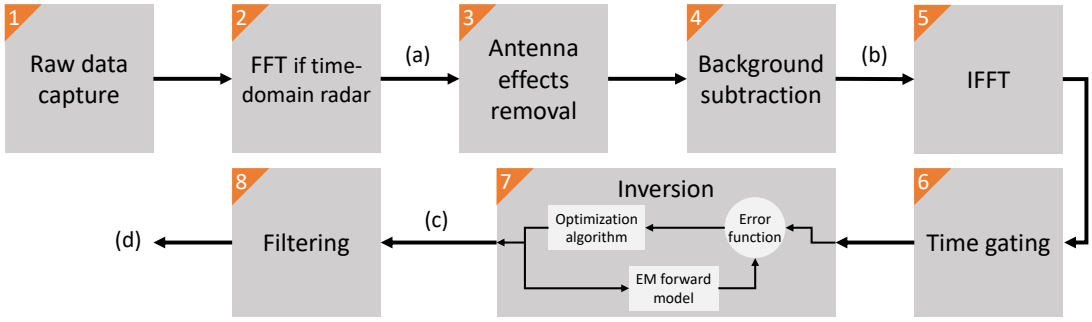


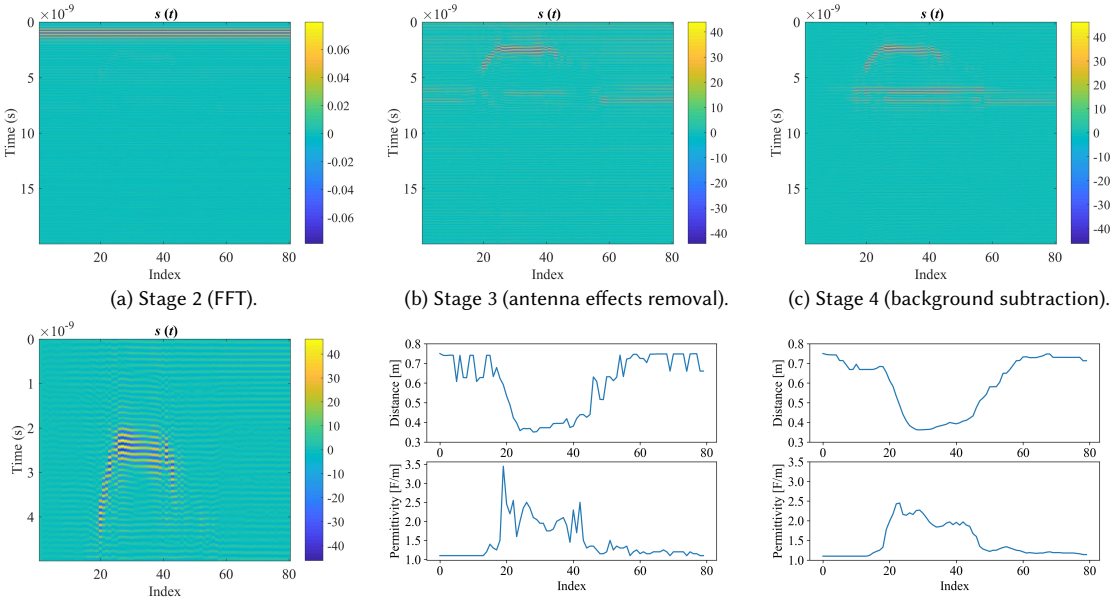
Fig. 2. The radar data processing pipeline. The recognition results of the best performing configurations are displayed for four significant stages of the pipeline: a (Fig. 8a, 9a, 10a, 11a), b (Fig. 8b, 9b, 10b, 11b), c (Fig. 8c, 9c, 10c, 11c), and d (Fig. 8d, 9d, 10d, 11d).

activities in a checkout counter of a convenience store and a typical office desk using a Walabot Pro sensor, the RF-radar version with 18 antennas that is capable of constructing a 3D image from the reflected radio waves. This sensor is deployed under the work surface and when the subject performs pre-defined activities, data is captured in the form of RF samples. Despite its apparent benefits, such as its low price ((~200$)), off-the-shelf availability, and small size, we are only aware of one paper using it for hand gesture recognition [56]. In this paper, Zhang *et al.* propose a deep neural network for continuous gesture recognition and evaluate it on a dataset of eight hand gestures (with 150 samples per gesture class) performed very close to the radar. The network is trained with 120 samples per gesture while the 30 remaining samples are used for testing. Thanks to the deep learning architecture and a large number of training templates, they reach a high accuracy at the price of a large number of templates for a limited set of gesture classes.

3 A NEW RADAR DATA PROCESSING PIPELINE FOR GESTURE RECOGNITION

We present a new radar data processing pipeline for hand gesture recognition that reduces the high dimensionality of raw radar data to only two physically meaningful features which are also independent of the radar itself. The pipeline is composed of eight stages, according to a principle-based approach, as follows (Fig. 2):

(1) **Raw data capture**. Raw data is captured from each radar antenna (Fig. 3a). Depending on the type of radar, the data may be in the time or frequency domain. For instance, the Walabot provides data in the time domain.
(2) **Fast Fourier Transform**. The radar signal is transformed from the time- to the frequency-domain if raw data are acquired in the time domain. This operation is required for the next stage, which must be performed in the frequency domain.
(3) **Removal of radar source and antenna effects**. The radar equation [29, 30] is applied to the raw frequency-domain signal to remove the radar source and antenna effects (*e.g.,* internal reflections and transmissions) and antenna-target interactions (Fig. 3b). By filtering out parasitic reflections, this stage enables the user's reflections to stand out from the rest of the signal. In addition, the radar data become normalized (Green's function) and thereby independent of the radar itself.
(4) **Removal of the background scene**. Using the superposition principle, the first frame of a gesture is subtracted from the radar signal to remove the remaining reflections from static reflectors, such as walls, furniture, or other



(a) Stage 2 (FFT). (b) Stage 3 (antenna effects removal). (c) Stage 4 (background subtraction).

(d) Stage 6 (time gating). (e) Stage 7 (inversion). (f) Stage 8 (filtering).

Fig. 3. Intermediate results of the processing of the "push with palm" gesture recorded with antenna pair 3 of the Walabot.

objects (Fig. 3c). This stage is required to ensure accurate feature extraction in the inversion stage, as reflections from other (static) objects could be confused with the user's hand.

(5) **Inverse Fast Fourier Transform**. The filtered radar signal is transformed from the frequency to the time domain to simplify the computations in the next stages.

(6) **Time gating**. The time-domain data is truncated to only keep the portion of the signal that is relevant for gesture recognition (Fig. 3d). Here, only the signal received within a given time window is kept. This removes useless information, such as objects that are too far away from the radar, thus improving accuracy and reducing the processing time of the next stages.

(7) **Inversion**. Two physically meaningful features are extracted from the filtered time-domain radar data by performing data inversion [30] (Fig. 3e). The first feature, the hand-radar distance, provides information about the hand trajectory. The second feature, the effective permittivity of the medium defined by the hand gives insights into the configuration of the hand (*e.g.*, whether the palm is facing towards the radar). The drastic data reduction enables the use of simple template-matching algorithms for gesture recognition. Reduced data from multiple antennas can be combined to improve recognition accuracy with a limited impact on recognition time.

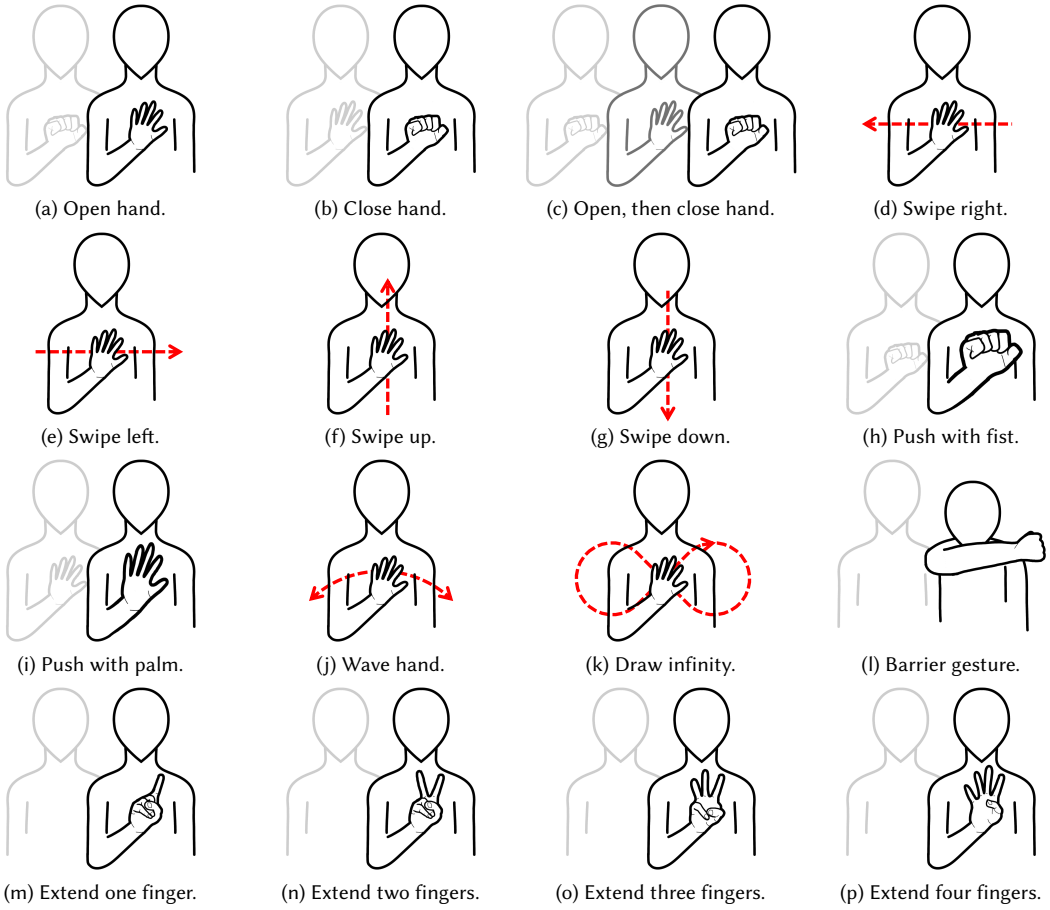
(8) **Filtering**. A moving-average filter with a window of length 5 is applied to smoothen the variations due to potential errors in the inversion process (Fig. 3f). In the absence of filtering, abrupt changes in estimated distance and/or permittivity value could negatively impact the accuracy of gesture recognizers. This step is in particular necessary for scenes for which the signal to noise ratio is poor (mainly depending on hand distance and orientation).

Fig. 3 shows the signal of the "push with palm" gesture recorded by one antenna pair of the Walabot throughout the main stages of the pipeline. Practitioners could benefit from using only part of the pipeline since it is flexible to use depending on constraints imposed by the target usage context. For instance, they could choose to use output data from

the FFT, background subtraction, inversion, or filtering stages. Fig. 2 provides, for stages 2, 4, 7, and 8 of the pipeline, the best recognition rate obtained in our evaluation (Section 5), as well as the corresponding execution time and average gesture size. The latter corresponds to the estimated space occupied by a gesture in RAM memory and depends on the duration of the gesture (80 frames in our setup), the number of selected antenna pairs, and the amount of data returned by the processing stage (*e.g.*, stages 7 and 8 return two 64-bit floating-point numbers for each antenna pair).

4 HAND GESTURE DATASETS

Prior to acquiring any dataset, we built a custom radar called the "horn" (Fig. 5, bottom right), equipped with an antenna system consisting of a linearly polarized, double-ridged broadband horn antenna (BBHA 9120 A, Schwarzbeck Mess-Elektronik, Germany). The antenna dimensions are 22 cm (8.66 in.) long and 14×24 cm^2 (5.51×9.45 $in.^2$) aperture area. The antenna nominal frequency range is 0.8 to 5 GHz, and its isotropic gain ranges from 6 to 18 dBi. The high directivity of the antenna (45° 3 dB beamwidth in the E-plane and 30° in the H-plane at 1 GHz) makes it suitable for our application to gesture recognition.



(a) Open hand. (b) Close hand. (c) Open, then close hand. (d) Swipe right.
(e) Swipe left. (f) Swipe up. (g) Swipe down. (h) Push with fist.
(i) Push with palm. (j) Wave hand. (k) Draw infinity. (l) Barrier gesture.
(m) Extend one finger. (n) Extend two fingers. (o) Extend three fingers. (p) Extend four fingers.

Fig. 4. The hand gesture set used in our study.

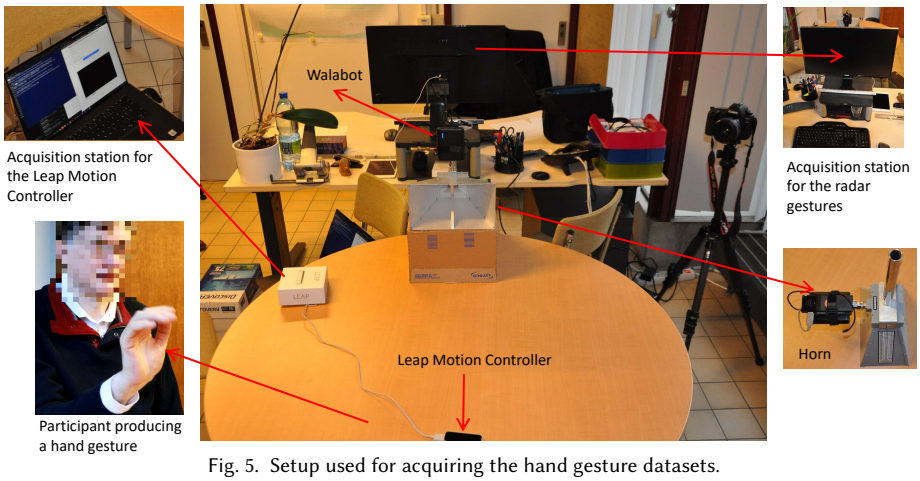


Fig. 5. Setup used for acquiring the hand gesture datasets.

Sixteen frequent gestures (Fig. 4) were determined by consolidating results obtained in [22, 36, 40] and in a gesture elicitation study [50] conducted with 30 participants (13 female, 17 male), aged between 18 and 58 years old, M=27.4, SD=12.6. Their occupations included language studies, computer science studies, engineering studies, journalism, law, and teaching. Consequently, a dataset of 16 gesture classes was recorded using three different sensors simultaneously to align their corresponding data (see Fig. 5 for the setup): the Walabot, the horn, and a Leap Motion Controller (LMC). Two versions of the dataset were recorded: (1) the "Walabot dataset", which consists of gestures recorded simultaneously with the Walabot and the LMC, and (2) the "Horn dataset" which consists of gestures recorded simultaneously with the horn antenna and the LMC. The use of an LMC and a horn antenna serves as a reference point for the performance of the Walabot. Five samples were acquired for each gesture from one right-handed user (male, 55 years old, without any prior experience with radar gestures). Although similar, the recording procedure slightly differs for both datasets:

(1) **Walabot dataset**. For each gesture, the time-domain radar data from six of the twelve antenna pairs are recorded in a file. 80 slow-time samples (*i.e.*, frames) are recorded per gesture at a rate of about 20 frames per second. The data is truncated to keep only the first 1024 fast-time samples out of 4096. This drastically reduces file size but also divides the maximum range by four. Simultaneously, data from the LMC is recorded using a custom tool[1].
(2) **Horn dataset**. For each gesture, the frequency-domain data from the horn antenna is recorded in a file at a rate of about 10 frames per second. Contrary to the Walabot dataset, the recording is stopped immediately after a gesture is completed, resulting in a variable number of frames per gesture. The LMC data are recorded simultaneously using our custom recorder.

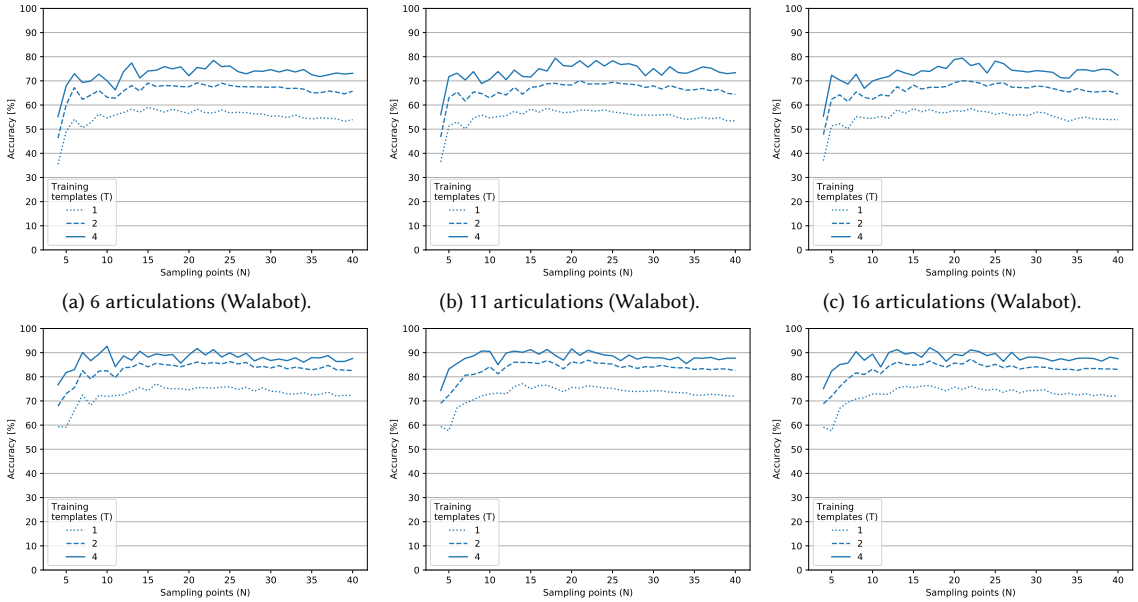
5 EVALUATION

After describing the general procedure for evaluation, we present the main results on the Horn and Walabot datasets.

5.1 General Procedure

All studies detailed in the rest of this section follow the same overall testing procedure. The *recognition rate* (*i.e.*, the ratio of correctly recognized gestures divided by the number of trials) and the *execution time* (*i.e.*, the time for recognizing the class of a candidate gesture) are computed for each combination of the studied variables (*e.g.*, the number of training

[1]The code of the recorder is available on GitHub at https://github.com/sluyters/LeapGesturePlayback.



(a) 6 articulations (Walabot). (b) 11 articulations (Walabot). (c) 16 articulations (Walabot).

(d) 6 articulations (Horn antenna). (e) 11 articulations (Horn antenna). (f) 16 articulations (Horn antenna).

Fig. 6. The accuracy of Jackknife with respect to the number of sampling points and training templates on the LMC data of the Walabot (top) and Horn antenna (bottom) dataset.

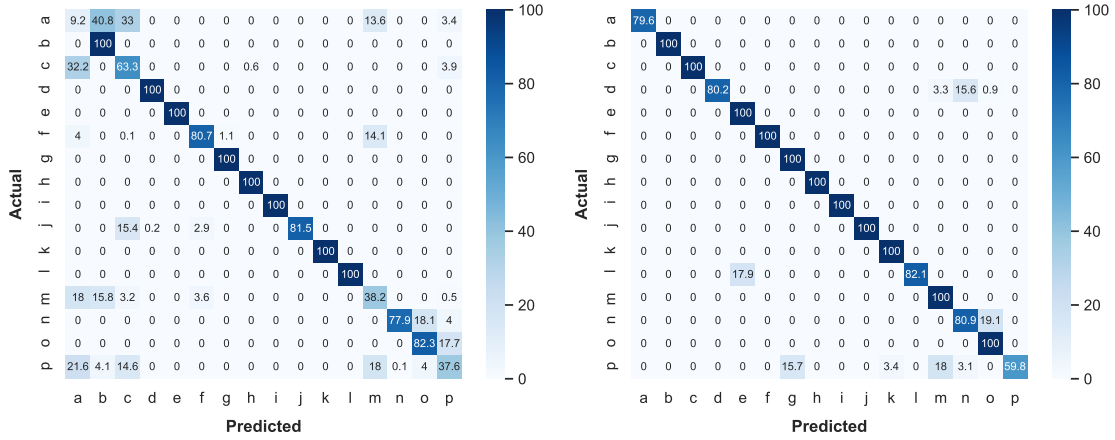
templates) in a user-dependent scenario [46], a typical procedure based on a one-leave out cross validation [57]. For each variable combination and for each gesture class, three steps are repeated 1000 times as follows: (1) select a random testing sample from the gesture set; (2) train the recognizer using a set of randomly selected samples, produced by the same user as the testing sample; and (3) recognize the testing sample. We consistently used the Jackknife recognizer [44] in all subsequent testings because this recognizer is efficient on multiple data types, being modality agnostic. All tests are run on a laptop with an Intel i7-10875H CPU and 32GB of DDR4 RAM running Windows 10.

5.2 Leap Motion Controller

In this first study, we evaluate the accuracy of Jackknife on the raw LMC data from the Walabot and horn antenna datasets following the procedure defined in Section 5.1. The study is within-factors with four independent variables:

(1) Dataset: nominal variable with 2 conditions, representing the datasets: Walabot dataset, Horn dataset.
(2) Number of Templates: numerical variable with 3 conditions, representing the number of templates per gesture class used to train the recognizer: T={1, 2, 4}.
(3) Number of Sampling Points: numerical variable with 5 values representing the number of points per gesture template: N={$x \in \mathbb{N} \mid 4 \leq x \leq 40$}.
(4) Number of Articulations: numerical variable with three conditions, representing the number of articulations of the hand taken into account by the recognizer: A∈{6, 11, 16}.

Each gesture sample consists of a sequence of frames, where each frame contains two elements: (1) a timestamp and (2) a vector of length $3 \times A$ which results from the concatenation of the 3D coordinates (x,y,z) from all selected articulations at that instant. Fig. 6 depicts the evolution of accuracy with respect to the number of sampling points on



(a) Walabot dataset. (b) Horn dataset.

Fig. 7. Normalized confusion matrices for the best configuration with data acquired from the LMC on the Walabot and Horn datasets. The values in each cell are represented as percentages.

the Horn and Walabot datasets. The recognition rate decreases sharply with $N<7$ but is mostly stable for higher values of N. For all configurations, the recognition rates on the Horn dataset are consistently higher than on the Walabot dataset, which may be explained by the slightly different recording procedure of both datasets (fixed *vs.* variable-length recordings, Section 4). The highest recognition rates are 79.4% (A=16, N=21) on the Walabot dataset and 92.7% (A=6, N=10) on the Horn dataset. Execution time is short at 0.7 and 0.2 ms, respectively, and stays under 1 ms in most configurations. Overall, the execution time increases with the number of sampling points, articulations, and training templates. Fig. 7 shows the confusion matrices for the best configurations of Jackknife [44] on the Walabot and Horn datasets. Smaller-scale hand gestures, including "extend one/two/three/four fingers" and "open/close hand" are less accurately recognized by the LMC in both the Walabot and Horn datasets. This could be caused by self-occlusion, where part of the hand (*e.g.*, the fingers) is occluded by another part of the hand (*e.g.*, the hand palm). Because the LMC was placed on a desk, under the participant's hands, it was not always possible to determine the exact position of the fingers, thus resulting in high confusion between these gestures. The confusion between the "open hand", "close hand", and "open, then close hand" gestures on the Walabot dataset seems caused by the fixed length of the recordings, which can pick up unintended motion after the relevant gesture has been performed. For instance, if a participant closes the hand shortly after performing the "open hand" gesture, the gesture could be interpreted as "open, then close hand". This does not happen on the Horn dataset as the recordings are stopped immediately after the completion of a gesture.

5.3 Walabot

We conduct two studies for the Walabot dataset, each with a specific objective. The first study investigates the best combinations of antenna pairs for recognizing our gesture set while the second one analyses the impact of various parameters on gesture recognition for the best performing combinations of antenna pairs determined in the first study. The first study is within-factors with three independent variables:

(1) Processing Stage: nominal variable with 4 conditions, representing the processing stage at which the training data has been taken: FFT, Background subtraction, Inversion, Filtering.

AP	FFT RR (*M, SD*) [%]	FFT ET [ms]	Background subtraction RR (*M, SD*) [%]	Background subtraction ET [ms]	Inversion RR (*M, SD*) [%]	Inversion ET [ms]	Filtering RR (*M, SD*) [%]	Filtering ET [ms]
1st best	73.8 (26.6)	2.2	74.3 (20.1)	2.2	58.1 (29.3)	0.2	57.8 (29.4)	0.2
2nd best	72.1 (21.1)	4.0	74.2 (18.3)	2.9	55.7 (31.1)	0.2	55.6 (31.3)	0.2
3rd best	72.1 (23.7)	4.6	72.7 (26.9)	3.6	55.6 (31.9)	0.2	55.3 (32.1)	0.2
4th best	71.8 (27.3)	3.3	72.4 (25.8)	3.5	55.1 (25.8)	0.1	54.8 (25.2)	0.1
2	48.8 (27.0)	0.8	51.9 (24.1)	0.8	36.8 (27.4)	0.1	37.8 (23.2)	0.1
3	47.7 (31.1)	0.8	48.0 (29.7)	0.9	25.8 (27.7)	0.1	25.6 (27.2)	0.1
4	49.1 (25.5)	0.8	54.3 (30.7)	0.8	25.3 (21.7)	0.1	25.9 (21.7)	0.1
6	44.6 (31.0)	0.8	46.3 (33.4)	0.9	19.3 (18.1)	0.1	18.9 (17.0)	0.1
7	51.7 (26.1)	0.8	49.3 (27.8)	0.9	23.2 (23.4)	0.1	23.4 (23.4)	0.1
10	56.6 (29.7)	0.8	52.9 (25.9)	0.8	22.0 (21.9)	0.1	21.9 (21.8)	0.1
2, 3, 4, 6, 7, 10	72.1 (23.7)	4.6	70.3 (26.2)	4.3	52.4 (29.6)	0.2	52.2 (29.4)	0.2

Table 1. Recognition rate and execution time on Walabot data at four stages of the pipeline for the four best performing sets of antenna pairs, all single antenna pairs, and the set of all antenna pairs. T=4 and N=16. AP = antenna pairs, RR = recognition rate, ET = execution time.

(2) Number of Templates: numerical variable with 3 conditions, representing the number of templates per gesture class used to train the recognizer: T={1, 2, 4}.
(3) Antenna Pairs: nominal variable with 63 conditions, representing all the possible combinations of antenna pairs (excluding the empty set): AP = {((2), (3), (4), (6), (7), (10), (2, 3), ..., (3, 4, 6, 7, 10), (2, 3, 4, 6, 7, 10)}.

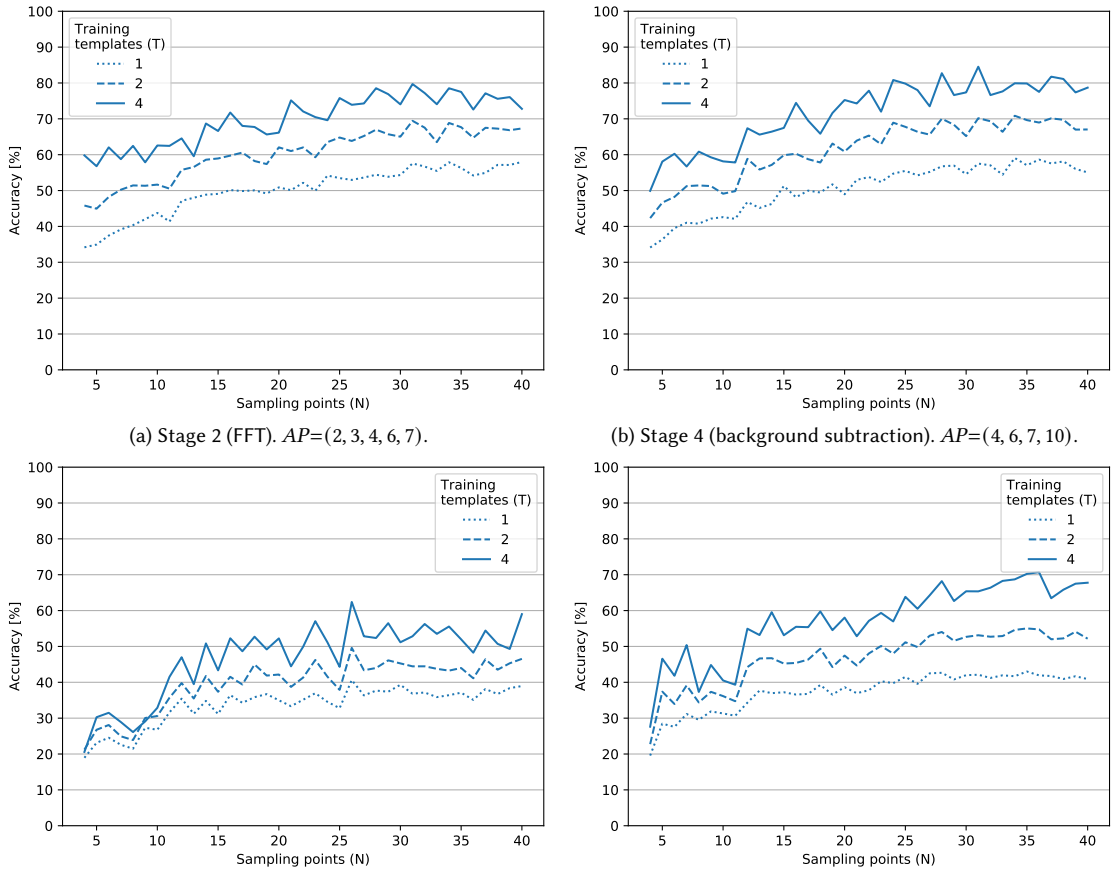
The number of sampling points is set to 16 in this first study. The second study is also within-factors, with four independent variables:

(1) Processing Stage: same as in the first study.
(2) Number of Templates: same as in the first study.
(3) Number of Sampling Points: numerical variable with 37 values representing the number of points per gesture template: N={$x \in \mathbb{N} \mid 4 \leq x \leq 40$}.
(4) Antenna Pairs: nominal variable with three conditions, representing the best performing combinations of antenna pairs determined in the first study. For instance, for processing stage "FFT", AP = {(4, 7, 10), (2, 3, 4, 6, 7), (2, 3, 4, 6, 7, 10)}.

In both studies, each gesture template consists of a sequence of frames, where each frame contains two elements: (1) a timestamp and (2) a vector with the radar data at that instant. For the inversion and filtering stage, the vector (of length $2 \times AP$) is the concatenation of the distance and permittivity values retrieved from all selected antenna pairs. For the other stages, the vector (of length $2 \times 34 \times AP$) results from the concatenation of the real and imaginary parts of the frequency-domain radar signal (34 frequencies) from all selected antenna pairs. In the rest of this section, we present the results from both studies at each processing stage.

5.3.1 FFT (Stage 2). The recognition rate differs widely depending on the set of antenna pairs used for gesture recognition (Table 1). Four different configurations (all with T=4 and N= 16) achieve >70% accuracy on the Walabot data before antenna effect removal and background subtraction: 73.8% with AP=(4, 7, 10), 72.1% with AP=(2, 3, 4, 6, 7), 72.1% with AP=(2, 3, 4, 6, 7, 10), and 71.8% with AP=(3, 4, 7, 10). Execution time increases with the number of antenna pairs and reaches up to 4.6 ms with 6 antenna pairs and T=4 training templates.

The three best-performing antenna pairs were selected for the second study. Fig. 8a plots the evolution of recognition rate with respect to the number of sampling points and the number of training templates for the best performing set



(a) Stage 2 (FFT). AP=(2, 3, 4, 6, 7).

(b) Stage 4 (background subtraction). AP=(4, 6, 7, 10).

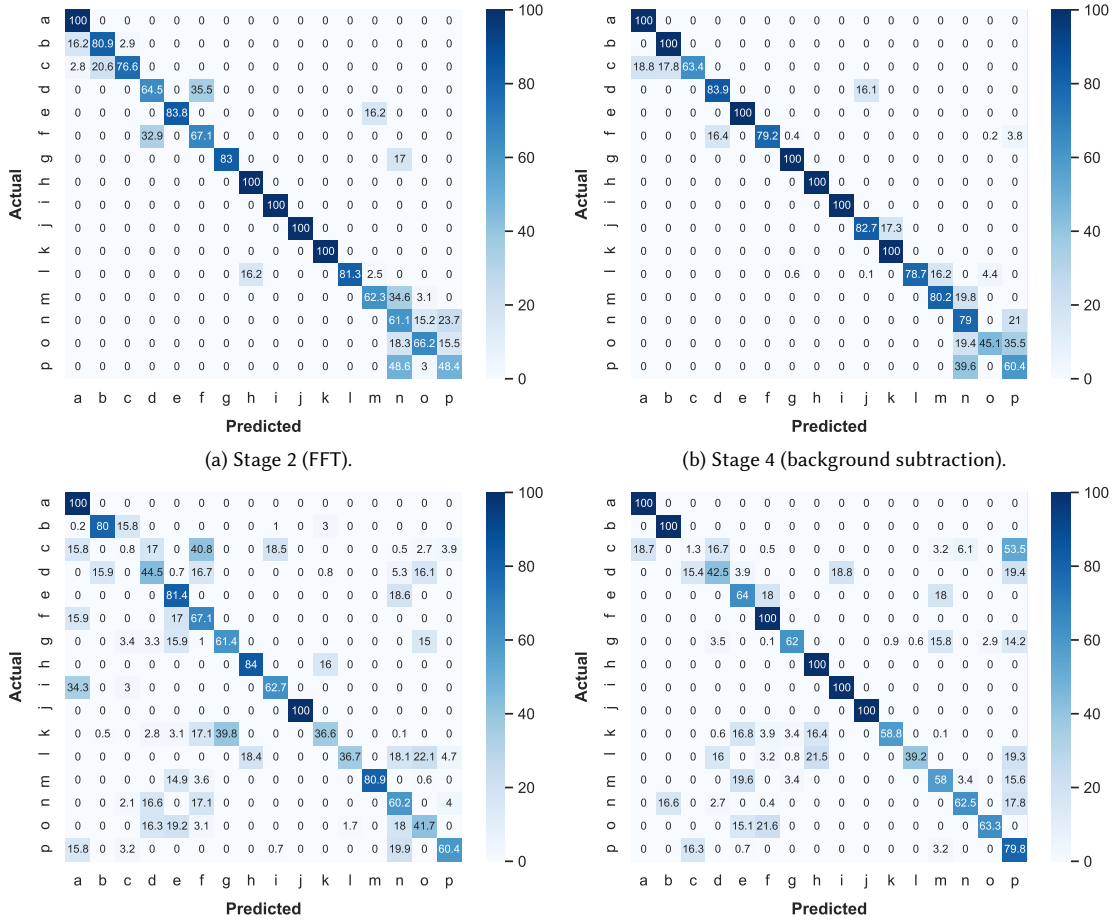
(c) Stage 7 (inversion). AP=(2, 3, 6, 7).

(d) Stage 8 (filtering). AP=(2, 3, 6, 7).

Fig. 8. The accuracy of Jackknife with respect to the number of sampling points and training templates for the best performing antenna pair with data acquired from the Walabot at four different stages of the pipeline.

of antenna pairs (AP=(2, 3, 4, 6, 7)). The recognition rate generally increases with the number of sampling points but varies more than on the LMC data. It reaches up to 79.7% (SD=16.9%) with T=4 and N=31, which is on par with the best performing configuration on the corresponding LMC data. The average execution time is 7.3 ms. The results on the two other sets of antenna pairs are similar with slightly lower recognition rates. Similar to the LMC data, Fig. 9a highlights some confusion between the "open/close hand" and "open, then close hand" gestures which could be due to the fixed length of the recordings (Section 5.2). There is also a lot of confusion between the "extend one/two/three/four fingers" gestures, which suggests that the Walabot might not be accurate enough for the recognition of fine-grained gestures performed at a medium distance from the sensor.

5.3.2 Background Subtraction (Stage 4). As for Stage 2, the accuracy differs widely depending on the set of antennas (Table 1). However, the overall results are better than with the raw data. 15 different configurations (all with T=4) achieve >70% accuracy, including: 74.3% with AP=(2, 6, 7), 74.2% with AP=(4, 6, 7, 10), 72.7% with AP=(2, 3, 4, 6, 7), and 72.4% with AP=(2, 4, 6, 7, 10). Execution time increases with the number of antenna pairs (up to 4.3 ms with 6 antenna



(a) Stage 2 (FFT).

(b) Stage 4 (background subtraction).

(c) Stage 7 (inversion).

(d) Stage 8 (filtering).

Fig. 9. Normalized confusion matrices for the best configuration with data acquired from the Walabot at four different stages of the pipeline. The values in each cell are represented as percentages.

pairs and 4 training templates). Compared to the raw data, the filtered radar data results in a higher recognition rate for four of the six individual antenna pairs.

The results from the second study show that recognition rates can reach more than 80% in some configurations, and up to 84.5% (SD=17.1%), with AP=(4, 6, 7, 10), T=4, and N=31. This represents an almost 5% gain compared to data from Stage 2. The average execution time for this configuration is 5.8 ms. The overall behavior of Jackknife is similar to the one observed on data from Stage 2. Fig. 9b shows similar results to the previous stage, with a noticeably higher recognition rate for most gesture classes. However, we still notice some confusion between fine-grained gestures such "extend one/two/three/four fingers", and between the "open/close hand" and "open, then close hand" gestures, which supports our conclusions from Section 5.3.1.

5.3.3 Inversion (Stage 7). The recognition rate for all sets of antenna pairs after the inversion stage is noticeably lower than after the first two stages (Table 1). Such a drop is expected, as the size of each frame is divided by 34 compared to

the first two processing stages. The best performing combinations of antenna pairs are: AP=(4, 6, 7) with 58.1% accuracy, AP=(2, 3, 6, 7) with 55.7% accuracy, AP=(2, 3, 6, 7, 10) with 55.6% accuracy, and AP=(2, 6, 7) with 55.1% accuracy.

As previously, the three best antenna pairs were selected for the second study. The recognition rate reaches up to 62.4% (SD=26.1%) with antenna pairs AP=(2, 3, 6, 7), T=4, and N=26. The execution time is very short, at about 0.4 ms. The recognition rate does not steadily increase with the number of sampling points and can change by more than 10% between two consecutive values of N. This behavior may be caused by errors in the inversion process, where the distance and apparent permittivity values corresponding to one frame are incorrectly estimated due to the low quality of the radar data (*e.g.,* if the received signal is barely above noise level). The filtering performed at Stage 8 should help alleviate this problem. Fig. 9c shows that the recognition rate is higher than 80% for only six gestures: "open hand", "close hand", "swipe left", "push with fist", "wave hand", and "extend one finger". The high accuracy for some of these gestures can be explained by the nature of the permittivity and distance metrics. The estimated permittivity increases when the hand opens (larger reflecting surface) and decreases when it closes (smaller reflecting surface), which enables to accurately distinguish between the first two gestures. The push with fist gesture is easy to differentiate from other gestures using the distance metric.

5.3.4 Filtering (Stage 8). Table 1 shows the recognition rate and execution time for some of the combinations of antenna pairs. The highest recognition rates are: 57.8% with AP=(4, 6, 7), 55.6% with AP=(2, 3, 6, 7), 55.3% with AP=(2, 3, 6, 7, 10), and 54.8% with AP=(2, 6, 7). All of these results were obtained with T=4 training templates.

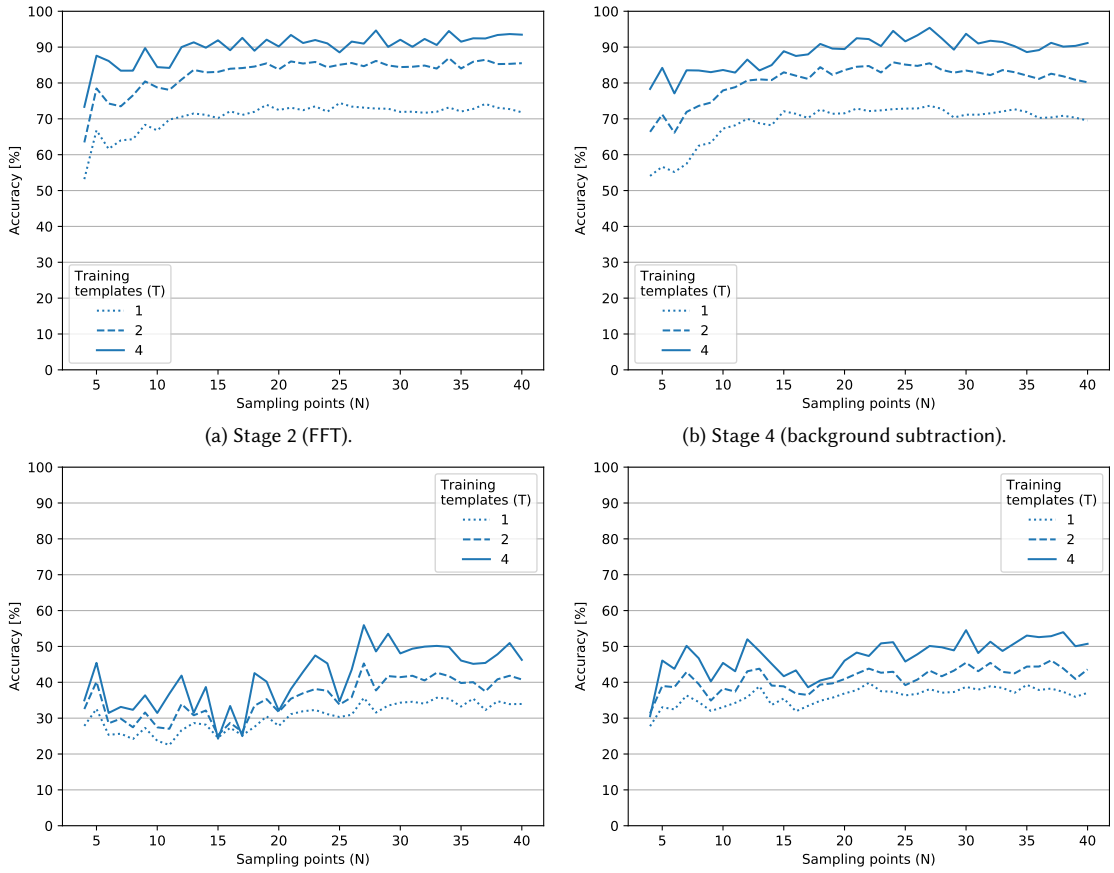
The three best antenna pairs were selected for the second study. Compared to the best configuration from Stage 7 (inversion), the recognition rate on the filtered data increases by about 8% to reach up to 70.7% (SD=28.7%) with T=4, N=36, and antenna pairs AP=(2, 3, 6, 7). The execution time is still extremely fast at 0.5 ms. Overall, the recognition rate seems to increase more steadily with the number of sampling points than in Stage 7. The confusion matrix in Fig. 9d shows that six gestures are now correctly recognized 100% of the time. However, the same behavior is visible, with poor recognition of fine-grained (*e.g.,* extend one/two/three/four fingers) and "swiping" gestures.

5.4 Horn Antenna

We conducted a study to investigate the impact of various parameters on gesture recognition. The study is within-factors with three independent variables:

(1) Processing Stage: nominal variable with 4 conditions, representing the processing stage at which the training data has been taken: FFT, Background subtraction, Inversion, Filtering.
(2) Number of Templates: numerical variable with 3 conditions, representing the number of templates per gesture class used to train the recognizer: T=$\{1, 2, 4\}$.
(3) Number of Sampling Points: numerical variable with 37 values representing the number of points per gesture template: N=$\{x \in \mathbb{N} \mid 4 \leq x \leq 40\}$.

Each gesture template consists of a sequence of frames, where each frame contains two elements: (1) a timestamp and (2) a vector containing the radar data at that instant. For the inversion and filtering stages, the vector is of length 2 and contains the distance and apparent permittivity evaluated at that instant. For the other stages, the vector is of length 2×89 and contains the real and imaginary parts of the frequency-domain radar signal (89 frequencies). Fig. 10 shows, for the four processing stages, the accuracy of Jackknife depending on the number of sampling points and training templates. In addition, Fig. 11 shows the confusion matrix of the best performing configuration for each stage.



(a) Stage 2 (FFT).

(b) Stage 4 (background subtraction).

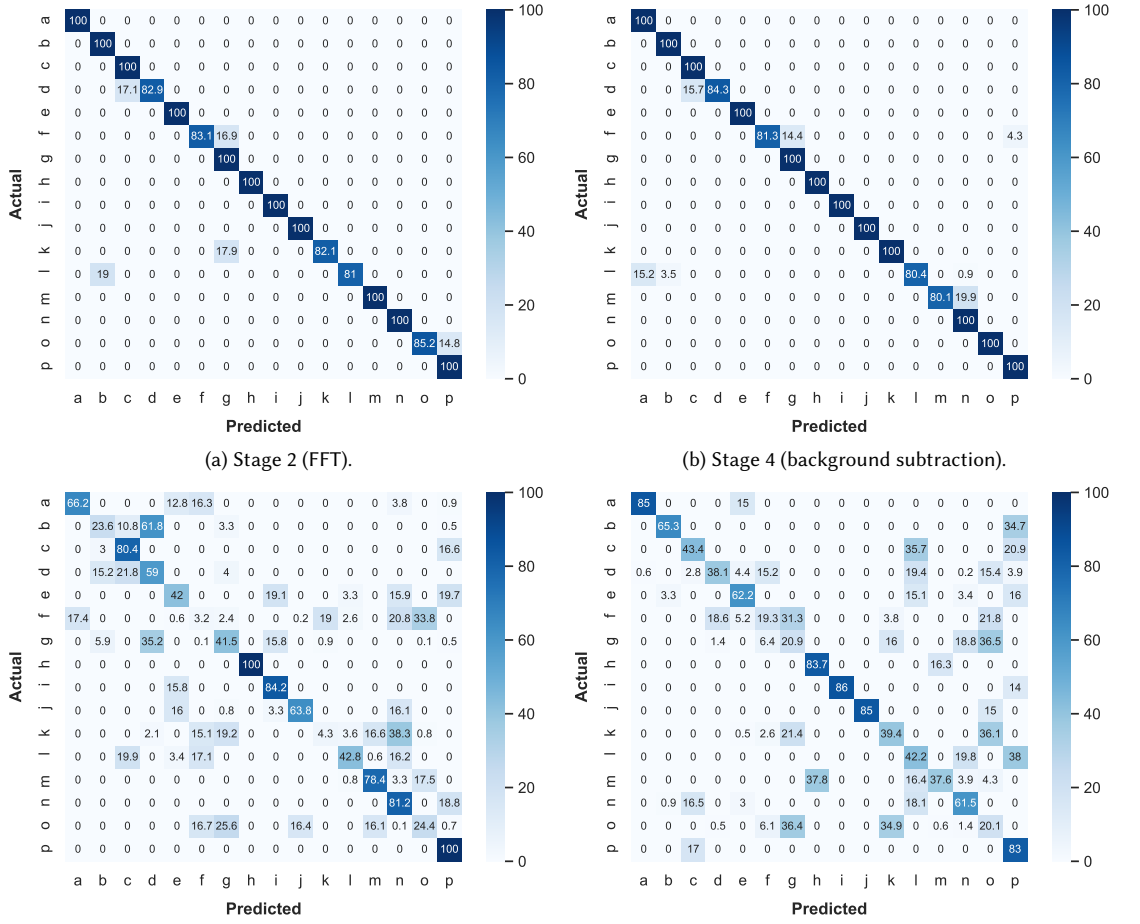
(c) Stage 7 (inversion).

(d) Stage 8 (filtering).

Fig. 10. The accuracy of Jackknife with respect to the number of sampling points and training templates with data acquired from the custom radar at four different stages of the pipeline.

5.4.1 FFT (Stage 2). With T=4 training templates, the recognition rate before antenna effects removal and background subtraction is already very high. Hovering around 90% with more than N=12 sampling points, it reaches up to 94.6% with T=4 and N=28 (Fig. 10a). In this configuration, the average execution time for recognizing a gesture is 3.8 ms. Both the accuracy and execution time increase with the number of training templates. From the confusion matrix (Fig. 11a), we see that 11 of the 16 gesture classes are correctly recognized 100% of the time. The five other gestures are "swipe right", "swipe up", "draw infinity", "barrier gesture", and "extend three fingers". The low recognition rate for these gestures may be due to the single horn radar antenna that is used, making it difficult to resolve motion performed in the plane orthogonal to the direction of the emitted radar signal. In that respect, an antenna array is preferable.

5.4.2 Background Subtraction (Stage 4). The removal of antenna effects and background subtraction enable even higher recognition rates, reaching 95.4% (SD=8.31%) with T=4 and N=27 (Fig. 10b). The resulting confusion matrix (Fig. 11b) is very similar to the confusion matrix obtained after the FFT, with some minor differences. 100% recognition rate is



(a) Stage 2 (FFT).

(b) Stage 4 (background subtraction).

(c) Stage 7 (inversion).

(d) Stage 8 (filtering).

Fig. 11. Normalized confusion matrices for the best configuration with data acquired from the custom radar at four different stages of the pipeline. The values in each cell are represented as percentages.

achieved for 12 of the 16 gesture classes, with the four other gestures being "swipe right", "swipe up", "barrier gesture", and "extend one finger".

5.4.3 Inversion (Stage 7). After the inversion stage, the recognition rate decreases significantly and reaches a maximum of 55.9% (SD=31.3%) with T=4 and N=27 (Fig. 10c). Average execution time is very short at 0.2 ms. Fig. 11c shows that the recognition rate is higher than 80% for only five gestures: "open, then close hand", "push with fist", "push with palm", "extend two fingers", and "extend four fingers". Similar to the Walabot dataset, the high accuracy for some of these gestures can be explained by the nature of the permittivity and distance metrics (see Section 5.3.3). The permittivity measure allows the distinction between the "open, then close hand" and the "extend two/four fingers" gestures. The "push with fist/palm" gestures are easy to differentiate from other gestures using the distance metric and the permittivity measure allows to distinguish between the two. As in the previous stages, "Swipe" gestures are not accurately recognized. In addition, the recognition rate varies considerably with the number of sampling points. For

instance, when T=4, the recognition rate goes from 33.4% with N=16 to 25.0% with N=17, and then goes up to 42.6% with N=18. As explained in Section 5.3.3, these variations may be caused by errors in the inversion process due to the relatively low signal to noise ratio for a series of scenes.

5.4.4 Filtering (Stage 8). As expected, the recognition rate is more stable with respect to the number of sampling points after filtering. However, it only reaches 54.5% (SD=25.0%) with T=4 and N=30 (Fig. 10d), which is lower than in the previous stage and thus still far too low for accurate gesture recognition. Execution time is still very short at 0.2 ms. The confusion matrix for this configuration (Fig. 11d) shows that the recognition rate exceeds 80% for only five gestures: "open hand", "push with fist", "push with palm", "wave hand", and "extend four fingers".

6 DISCUSSION AND IMPLICATIONS FOR DESIGN

This section discusses the results of the testing presented in the previous section and their implications for the design of radar-based gesture interaction.

- **Favor gestures with a highly differentiable surface of exposure.** For instance, the "push with palm" and "push with fist" gestures were accurately recognized in most configurations because the palm has a larger exposure surface than the fist, resulting in a larger amplitude of the received radar signal. This difference in amplitude can be used by a recognizer to differentiate between the two gestures. On the other hand, gestures such as "extend one/two/three/four fingers" were less accurately recognized (Section 5.3.1) because their surface of exposure was similar. Future work could study the impact of reducing the hand-radar distance on the ability to distinguish between gestures with a similar surface of exposure.
- **Favor gestures with motion parallel to the radar beam,** such as the "push with palm" and "push with fist" gestures. Indeed, as the angular resolution of our radars is lower than their range resolution, it is easier to recognize gestures performed parallel to the radar beam. On the other hand, gestures performed in the plane orthogonal to the radar beam, such as the "swipe left/right/up/down" gestures are more difficult to differentiate, resulting in lower accuracy (see for example Section 5.4.1). Future work may investigate whether techniques such as relying on a larger set of antenna pairs or placing multiple radar sensors around the interaction space could circumvent this limitation of radar sensors.
- **Keep a minimum of four templates.** Our testing showed that the accuracy greatly improved when increasing the number of training templates from 1 to 4 (see for example Section 5.4.2). Future work could evaluate the impact of using more than 4 training templates on accuracy and execution time. However, keeping the number of training templates small is beneficial to the end-users, as it allows them to define their own gestures with minimal time and effort, by recording only a small number of templates.
- **Favor the best combinations of antenna pairs.** Our testing revealed that accuracy varies depending on the selected antenna pairs of the Walabot (Table 1), and that the best results were obtained with sets of three or more pairs. The best performing antenna pairs for each stage are: $AP = (2, 3, 4, 6, 7)$ for Stage 2 (Section 5.3.1), $AP = (4, 6, 7, 10)$ for Stage 4 (Section 5.3.2), $AP = (2, 3, 6, 7)$ for Stage 7 (Section 5.3.3), and $AP = (2, 3, 6, 7)$ for Stage 8 (Section 5.3.4). The sets $AP = (4, 6, 7, 10)$ and $AP = (2, 3, 6, 7)$ consist of only non-redundant antenna pairs, which may explain their higher performance compared to other sets of antenna pairs. Redundant antennas pairs rely on the same two antennas. For instance, pairs 1 and 7 are redundant as they both rely on antennas 1 and 4 (see Fig. 1).

7 LIMITATIONS

We proposed and tested a radar data processing pipeline for hand gesture recognition. This section discusses the limitations of this work, which are divided into three categories: limitations of the Walabot, limitations of the proposed pre-processing strategy, and limitations of the evaluation procedure.

Walabot-induced limitations. The bandwidth of the Walabot EU/CE version used in this paper is narrower than the US/FCC version, resulting in a lower resolution. However, our pre-processing strategy should support the US/FCC version with only minor changes. In addition, while the Walabot features 18 antennas, three of them cannot be selected (Fig. 1) and the set of available antenna pairs is limited. For instance, it is not possible to combine relevant pairs from the first and second profiles. Compared to more expensive radar systems, the Walabot suffers from a relatively low signal-to-noise ratio and interferences between the emitter and receiver antennas, which hinder the ability to discern hand echoes from background noise and interferences. While the radar-antenna effects removal and background subtraction stages described in Section 3 help clean up the data, the relatively low signal-to-noise ratio with respect to the requirements of the ad hoc application remains problematic, especially for the inversion stage (Section 5.3.3).

Limitations of the proposed data pre-processing pipeline. The removal of antenna effects requires calibration to construct the radar antenna model. The antenna calibration process requires taking a series of radar measurements at various distances from a large smooth metallic surface, which is impractical without the proper equipment. However, calibration could be performed only once (*e.g.,* in the factory) if the Walabot is shown to be stable over time and with respect to operation temperature changes. While the proposed dimension reduction performed at the inversion stage is physically meaningful, it results in a loss of information compared to raw radar data, which can negatively impact gesture recognition accuracy (see Table 1). The inversion process can also fail when the signal-to-noise ratio is low and thus result in incorrect distance and permittivity estimates (main identified limitation) (Section 5.3.3). In addition, the estimated relative permittivity depends on properties of the user's hand, such as shape, size, and whether the user is wearing gloves. For a gesture set to be compatible with most users, its gesture recordings should thus cover a wide range of hand types. Another limitation is that the inversion process may fail or return erroneous values if more than one hand and/or more than one user is present at a short distance from the sensor. A potential, but computationally-intensive solution would be to set up the inversion process so that it returns one (distance, permittivity) pair for each hand. Finally, the present full-wave inversion process is slow, preventing the recognition of gestures in real-time. A drastic reduction in execution time is however possible and will be the subject of future works.

Limitations of the evaluation. Although our testing dataset contains a large number of gestures classes, only five samples were collected from one participant for each of its gestures. Further testing with data from more participants would give us insights into the transferability of gestures across participants. Another limitation is that we only tested one-handed gestures recorded with one user in the field of view of the sensors. We believe that the relatively narrow field of view and short range of our sensors makes multi-user interactions impractical. However, it is worth investigating the impact of the number of users on gesture recognition accuracy, both on frequency data and on data from the inversion stage. In addition, our evaluation covers only six out of the 12 antenna pairs proposed in profile 2. Testing more antenna pairs, including the 40 pairs from profile 1 would allow us to identify the best sets of pairs for accurate and efficient gesture recognition. Finally, slight differences in the recording procedures of the Walabot and Horn datasets (see Section 4) may affect our ability to compare results across the two sensors. While both datasets were also recorded using the LMC, which provides a good reference point, recording the two datasets using the exact same procedure would further strengthen the validity of our results.

8 CONCLUSION

The use of microwave radar sensors in gesture recognition has several key advantages compared to vision-based sensors, including providing the user with an enhanced feeling of privacy and being less sensitive to ambient conditions (*e.g.,* lighting and weather). However, due to the complexity of raw radar signals, most existing works rely on advanced machine learning techniques for radar gesture recognition.

In this paper, we proposed a new pipeline for radar gesture recognition that filters the raw radar signals and reduces the dimension of radar signal to only two physically meaningful features: the hand-radar distance and the effective permittivity of the medium defined by the hand. We thereby combine purely physical electromagnetic modeling and artificial intelligence to recognize gestures. To evaluate the pipeline, we recorded 16 gestures with three different sensors: an off-the-shelf radar (the Walabot), a custom-built radar (the "horn"), and a cheap vision-based sensor (the LMC). We then tested the accuracy of Jackknife, a popular template-matching recognizer, on the gesture set at various stages of the processing pipeline. We observed that removing the antenna effects and subtracting the background scene generally enabled higher accuracy by filtering out irrelevant reflections from the signal. The inversion stage greatly reduced the size of the gestures and the average execution time at the expense of accuracy. However, while accuracy was too low for practical applications when using only one antenna pair, using larger sets of antenna pairs seemed to yield a significant increase in accuracy.

In future work, we will evaluate the performance of other radar configurations. For instance, we could capture gestures using all 40 antenna pairs of the Walabot, using the Walabot US/FCC instead of the EU/CE version, or using multiple radars antennas at placed various positions around the interaction space. We also plan on investigating whether a dataset recorded in one environment (*e.g.,* a cluttered room) and processed by our pipeline could be used to train Jackknife for recognizing gestures in a different environment (*e.g.,* outdoors). Similarly, we could investigate whether our pipeline allows a gesture produced by one radar to be recognized using a dataset recorded with a different radar, thus enabling the recognition of a user's gestures using the same training dataset, independently of the radar sensor that they are using. Finally, we will expand our gesture set with recordings from more users. This will enable more extensive testing, including evaluating the performance of our pipeline on recordings produced by a different user than the training data.

OPEN SCIENCE

Our website https://sites.uclouvain.be/ingenious/projects/walabot-hgr provides the reader with useful resources, including the Horn and Walabot datasets (Section 4), and the data from the testing (Section 5).

ACKNOWLEDGMENTS

The authors of this paper are very grateful to the anonymous IUI (meta-)reviewers whose suggestions helped improve and clarify this manuscript. They also thank the participants of the gesture study reported in the paper for their involvement. The authors acknowledge funding received by Wallonie-Bruxelles-International (WBI), Belgium, under grant SUB/2021/519018 and UEFISCDI, Romania, under grant PN-III-CEI-BIM-PBE-2020-0001/1BM/2021 (Project "RadarSense"). Arthur Sluÿters is funded by the "Fonds de la Recherche Scientifique - FNRS" under Grant n°40001931.

REFERENCES

[1] G. Agresti and S. Milani. 2019. Material Identification Using RF Sensors and Convolutional Neural Networks. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '19)*. 3662–3666. https://doi.org/10.1109/ICASSP.2019.8682296

[2] Shahzad Ahmed, Dingyang Wang, Junyoung Park, and Sung Ho Cho. 2021. UWB-gestures, a public dataset of dynamic hand gestures acquired using impulse radar sensors. *Scientific Data* 8, 102 (April 2021). https://doi.org/10.1038/s41597-021-00876-0

[3] Ahmad Akl and Shahrokh Valaee. 2010. Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2270 – 2273. https://doi.org/10.1109/ICASSP.2010.5495895

[4] Mohammed Alloulah, Anton Isopoussu, and Fahim Kawsar. 2018. On Indoor Human Sensing Using Commodity Radar. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1331–1336. https://doi.org/10.1145/3267305.3274180

[5] F. Argelaguet, M. Ducoffe, A. Lécuyer, and R. Gribonval. 2017. Spatial and rotation invariant 3D gesture recognition based on sparse representation. In *Proc. of IEEE Symposium on 3D User Interfaces* (Los Angeles, CA, USA) *((3DUI '17))*. 158–167. https://doi.org/10.1109/3DUI.2017.7893333

[6] Daniel Ashbrook and Thad Starner. 2010. MAGIC: A Motion Gesture Design Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2159–2168. https://doi.org/10.1145/1753326.1753653

[7] Nuwan T. Attygalle, Luis A. Leiva, Matjaz Kljun, Christian Sandor, Alexander Plopski, Hirokazu Kato, and Klen Copic Pucihar. 2021. No Interface, No Problem: Gesture Recognition on Physical Objects Using Radar Sensing. *Sensors* 21, 17 (2021), 5771. https://doi.org/10.3390/s21175771

[8] Daniel Avrahami, Mitesh Patel, Yusuke Yamaura, and Sven Kratz. 2018. Below the Surface: Unobtrusive Activity Recognition for Work Surfaces Using RF-Radar Sensing. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 439–451. https://doi.org/10.1145/3172944.3172962

[9] Daniel Avrahami, Mitesh Patel, Yusuke Yamaura, Sven Kratz, and Matthew Cooper. 2019. Unobtrusive Activity Recognition and Position Estimation for Work Surfaces Using RF-Radar Sensing. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 11 (Aug. 2019), 28 pages. https://doi.org/10.1145/3241383

[10] A. D. Berenguer, M. C. Oveneke, H. Khalid, M. Alioscha-Perez, A. Bourdoux, and H. Sahli. 2019. GestureVLAD: Combining Unsupervised Features Representation and Spatio-Temporal Aggregation for Doppler-Radar Gesture Recognition. *IEEE Access* 7 (2019), 137122–137135. https://doi.org/10.1109/ACCESS.2019.2942305

[11] Andrew Bragdon, Rob DeLine, Ken Hinckley, and Meredith Ringel Morris. 2011. Code Space: Touch + Air Gesture Hybrid Interactions for Supporting Developer Meetings. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Kobe, Japan) *(ITS '11)*. Association for Computing Machinery, New York, NY, USA, 212–221. https://doi.org/10.1145/2076354.2076393

[12] Sanders Brandon. 2014. *Mastering Leap Motion.* Packt Publishing, Birmingham.

[13] R. Brunelli. 2009. *Template Matching Techniques in Computer Vision: Theory and Practice.* John Wiley & Sons, New York.

[14] Linfeng Chen, V. V. Varadan, C. K. Ong, and Chye Poh Neo. 2004. *Microwave Electronics: Measurement and Materials Characterization.* Wiley & Sons, New York, NY, USA.

[15] Hong Cheng, Lu Yang, and Zicheng Liu. 2016. Survey on 3D Hand Gesture Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 9 (2016), 1659–1673. https://doi.org/10.1109/TCSVT.2015.2469551

[16] Jae-Woo Choi, Si-Jung Ryu, and Jong-Hwan Kim. 2019. Short-Range Radar Based Real-Time Hand Gesture Recognition Using LSTM Encoder. *IEEE Access* 7 (2019), 33610–33618. https://doi.org/10.1109/ACCESS.2019.2903586

[17] Alex Colgan. 2017. How Does the Leap Motion Controller Work? https://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work/

[18] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat. 2017. 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset: SHREC'17 Track. In *Proceedings of the Workshop on 3D Object Retrieval* (Lyon, France) *(3Dor '17)*. Eurographics Association, Goslar, DEU, 33–38. https://doi.org/10.2312/3dor.20171049

[19] Zak Flintoff, Bruno Johnston, and Minas Liarokapis. 2018. Single-Grasp, Model-Free Object Classification using a Hyper-Adaptive Hand, Google Soli, and Tactile Sensors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1943–1950. https://doi.org/10.1109/IROS.2018.8594166

[20] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 623–635. https://doi.org/10.1145/3196709.3196741

[21] Daniela Grijincu, Miguel A. Nacenta, and Per Ola Kristensson. 2014. User-Defined Interface Gestures: Dataset and Analysis. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (Dresden, Germany) *(ITS '14)*. Association for Computing Machinery, New York, NY, USA, 25–34. https://doi.org/10.1145/2669485.2669511

[22] Eiji Hayashi, Jaime Lien, Nicholas Gillian, Leonardo Giusti, Dave Weber, Jin Yamanaka, Lauren Bedal, and Ivan Poupyrev. 2021. RadarNet: Efficient Gesture Recognition Technique Utilizing a Miniature Radar Sensor. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 5, 14 pages. https://doi.org/10.1145/3411764.3445367

[23] S. Hazra and A. Santra. 2019. Short-Range Radar-Based Gesture Recognition System Using 3D CNN With Triplet Loss. *IEEE Access* 7 (2019), 125623–125633. https://doi.org/10.1109/ACCESS.2019.2938725

[24] M. Hoffman, P. Varcholik, and J. J. LaViola. 2010. Breaking the status quo: Improving 3D gesture recognition with spatially convenient input devices. In *2010 IEEE Virtual Reality Conference (VR)*. 59–66.

[25] Jinmiao Huang, Prakhar Jaiswal, and Rahul Rai. 2019. Gesture-based system for next generation natural and intuitive interfaces. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 33, 1 (2019), 54–68. https://doi.org/10.1017/S0890060418000045

[26] Cloe Huesser, Simon Schubiger, and Arzu Çöltekin. 2021. Gesture Interaction in Virtual Reality. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 151–160.

[27] D. Kohlsdorf, T. Starner, and D. Ashbrook. 2011. MAGIC 2.0: A web tool for false positive prediction and prevention for gesture recognition systems. In *Face and Gesture 2011*. 1–6.

[28] Panayiotis Koutsabasis and Panagiotis Vogiatzidakis. 2019. Empirical Research in Mid-Air Interaction: A Systematic Review. *International Journal of Human–Computer Interaction* 35, 18 (2019), 1747–1768. https://doi.org/10.1080/10447318.2019.1572352

[29] Sébastien Lambot and Frédéric André. 2014. Full-Wave Modeling of Near-Field Radar Data for Planar Layered Media Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing* 52, 5 (2014), 2295–2303. https://doi.org/10.1109/TGRS.2013.2259243

[30] Sébastien Lambot, E.C. Slob, Idesbald van den Bosch, Benoit Stockbroeckx, and Marnik Vanclooster. 2004. Modeling of ground-penetrating Radar for accurate characterization of subsurface electric properties. *IEEE Transactions on Geoscience and Remote Sensing* 42, 11 (2004), 2555–2568. https://doi.org/10.1109/TGRS.2004.834800

[31] Gierad Laput and Chris Harrison. 2019. *Sensing Fine-Grained Hand Activity with Smartwatches.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300568

[32] Joseph J. LaViola. 2013. 3D Gestural Interaction: The State of the Field. *International Scholarly Research Notices* 2013, Article 514641 (2013), 18 pages. https://doi.org/10.1155/2013/514641

[33] Feifei Li, Yujun Li, Baozhen Du, Hongji Xu, Hailiang Xiong, and Min Chen. 2019. A Gesture Interaction System Based on Improved Finger Feature and WE-KNN. In *Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence* (Chegndu, China) *(ICMAI 2019)*. Association for Computing Machinery, New York, NY, USA, 39–43. https://doi.org/10.1145/3325730.3325759

[34] G. Li, S. Zhang, F. Fioranelli, and H. Griffiths. 2018. Effect of sparsity-aware time–frequency analysis on dynamic hand gesture classification with radar micro-Doppler signatures. *IET Radar, Sonar Navigation* 12, 8 (2018), 815–820. https://doi.org/10.1049/iet-rsn.2017.0570

[35] Tianxing Li, Xi Xiong, Yifei Xie, George Hito, Xing-Dong Yang, and Xia Zhou. 2017. Reconstructing Hand Poses Using Visible Light. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 71 (sep 2017), 20 pages. https://doi.org/10.1145/3130937

[36] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.* 35, 4, Article 142 (July 2016), 19 pages. https://doi.org/10.1145/2897824.2925953

[37] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. UWave: Accelerometer-Based Personalized Gesture Recognition and Its Applications. *Pervasive Mob. Comput.* 5, 6 (Dec. 2009), 657–675. https://doi.org/10.1016/j.pmcj.2009.07.007

[38] Mehran Maghoumi and Joseph J. LaViola. 2019. DeepGRU: Deep Gesture Recognition Utility. In *Advances in Visual Computing*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu (Eds.). Springer International Publishing, Cham, 16–31.

[39] Antigoni Mezari and Ilias Maglogiannis. 2017. Gesture Recognition Using Symbolic Aggregate Approximation and Dynamic Time Warping on Motion Data. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 342–347. https://doi.org/10.1145/3154862.3154927

[40] Sameera Palipana, Dariush Salami, Luis A. Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 27:1–27:27. https://doi.org/10.1145/3448110

[41] Joseph Paradiso, Craig Abler, Kai-yuh Hsiao, and Matthew Reynolds. 1997. The Magic Carpet: Physical Sensing for Immersive Environments. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia) *(CHI EA '97)*. Association for Computing Machinery, New York, NY, USA, 277–278. https://doi.org/10.1145/1120212.1120391

[42] Yuwei Ren, Jiuyuan Lu, Andrian Beletchi, Yin Huang, Ilia Karmanov, Daniel Fontijne, Chirag Patel, and Hao Xu. 2021. Hand gesture recognition using 802.11ad mmWave sensor in the mobile device. In *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. 1–6. https://doi.org/10.1109/WCNCW49093.2021.9419978

[43] Bert Schiettecatte and Jean Vanderdonckt. 2008. AudioCubes: a distributed cube tangible interface based on interaction range for sound design. In *Proceedings of the 2nd International Conference on Tangible and Embedded Interaction 2008, Bonn, Germany, February 18-20, 2008*, Albrecht Schmidt, Hans Gellersen, Elise van den Hoven, Ali Mazalek, Paul Holleis, and Nicolas Villar (Eds.). ACM, 3–10. https://doi.org/10.1145/1347390.1347394

[44] Eugene M. Taranta II, Amirreza Samiei, Mehran Maghoumi, Pooya Khaloo, Corey R. Pittman, and Joseph J. LaViola Jr. 2017. Jackknife: A Reliable Recognizer with Few Samples and Many Modalities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 5850–5861. https://doi.org/10.1145/3025453.3026002

[45] Laurens van der Maaten, Eric Postma, and H. Herik. 2009. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research* 10 (1 2009), 66–71. https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf

[46] Radu-Daniel Vatavu. 2013. The impact of motion dimensionality and bit cardinality on the design of 3D gesture recognizers. *International Journal of Human-Computer Studies* 71, 4 (2013), 387 – 409. https://doi.org/10.1016/j.ijhcs.2012.11.005

[47] P. Wang, J. Lin, F. Wang, J. Xiu, Y. Lin, N. Yan, and H. Xu. 2020. A Gesture Air-Writing Tracking Method that Uses 24 GHz SIMO Radar SoC. *IEEE Access* 8 (2020), 152728–152741. https://doi.org/10.1109/ACCESS.2020.3017869

[48] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 851–860. https://doi.org/10.1145/2984511.2984565

[49] Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Denis Fisseler. 2013. Analysis of the Accuracy and Robustness of the Leap Motion Controller. *Sensors* 13 (05 2013), 6380–6393. https://doi.org/10.3390/s130506380

[50] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. [n. d.]. User-defined Gestures for Surface Computing *(CHI '09)*. ACM, 1083–1092. https://doi.org/10.1145/1518701.1518866 ZSCC: 0001254.

[51] Mais Yasen and Shaidah Jusoh. 2019. A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science* 5 (Sept. 2019), e218. https://doi.org/10.7717/peerj-cs.218

[52] Hui-Shyong Yeo, Gergely Flamich, Patrick Schrempf, David Harris-Birtill, and Aaron Quigley. 2016. RadarCat: Radar Categorization for Input & Interaction. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 833–841. https://doi.org/10.1145/2984511.2984515

[53] Hui-Shyong Yeo, Ryosuke Minami, Kirill Rodriguez, George Shaker, and Aaron Quigley. 2018. Exploring Tangible Interactions with Radar Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 200 (Dec. 2018), 25 pages. https://doi.org/10.1145/3287078

[54] Hui-Shyong Yeo and Aaron Quigley. 2017. Radar Sensing in Human-Computer Interaction. *Interactions* 25, 1 (Dec. 2017), 70–73. https://doi.org/10.1145/3159651

[55] Wei Zeng, Cong Wang, and Qinghui Wang. 2018. Hand Gesture Recognition Using Leap Motion via Deterministic Learning. *Multimedia Tools Appl.* 77, 21 (Nov. 2018), 28185–28206.

[56] Bo Zhang, Lei Zhang, Mojun Wu, and Yan Wang. 2021. Dynamic Gesture Recognition Based on RF Sensor and AE-LSTM Neural Network. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1–5. https://doi.org/10.1109/ISCAS51556.2021.9401065

[57] F. Zhou, X. Li, and Z. Wang. 2019. Efficiently User-Independent Ultrasonic-Based Gesture Recognition Algorithm. In *2019 IEEE SENSORS*. 1–4. https://doi.org/10.1109/SENSORS43011.2019.8956774

[58] Shangyue Zhu, Junhong Xu, Hanqing Guo, Qiwei Liu, Shaoen Wu, and Honggang Wang. 2018. Indoor Human Activity Recognition Based on Ambient Radar with Signal Processing and Machine Learning. In *2018 IEEE International Conference on Communications (ICC)*. 1–6. https://doi.org/10.1109/ICC.2018.8422107