Leveraging Directional Antenna Capabilities for Fine-Grained Gesture Recognition

Pedro Melgarejo, Xinyu Zhang and Parameswaran Ramanathan
University of Wisconsin-Madison
{pmelgarejo, xyzhang, parmesh} @ece.wisc.edu

David Chu
Microsoft Research
davidchu@microsoft.com

ABSTRACT

This paper presents a recognition scheme for fine-grain gestures. The scheme leverages directional antenna and short-range wireless propagation properties to recognize a vocabulary of action-oriented gestures from the American Sign Language. Since the scheme only relies on commonly available wireless features such as Received Signal Strength (RSS), signal phase differences, and frequency subband selection, it is readily deployable on commercial-off-the-shelf IEEE 802.11 devices. We have implemented the proposed scheme and evaluated it in two potential application scenarios: gesture-based electronic activation from wheelchair and gesture-based control of car infotainment system. The results show that the proposed scheme can correctly identify and classify up to 25 fine-grain gestures with an average accuracy of 92% for the first application scenario and 84% for the second scenario.

Author Keywords

Wireless; Gestures Recognition.

ACM Classification Keywords

H5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION

The prevalence of wireless infrastructure and sensor-rich mobile devices is gradually enabling more intuitive communication between human and ambient objects, thus more closely resembling human-human communication. In particular, recent advances in machine learning and human-computer interaction have vastly taken advantage of the wireless and mobile devices, to enable many innovative technologies that can sense human hand/body gestures as control commands. Gesture recognition has found a diverse set of applications, e.g., 3D in-air user-interface for mobile and desktop computers [18], remote control of home appliances [15], sterilized operation of medical devices and distraction-free management of in-car infotainment system [26].

Behind these technologies lies a wide range of sensing interfaces. Classical gesture recognition solutions rely on computer vision [16, 23], i.e., processing of recorded video frames to identify the pattern. Vision-based approaches tend to be disturbed by irrelevant background patterns and low-light conditions. The infrared spectrum of sunlight often interferes with systems that rely on infrared, e.g., LeapMotion [9] and Kinect [17]. Hence, such systems are not intended for outdoor use. Furthermore, since video cameras may leak sensitive private information, people often consider them to be obtrusive. Alternative solutions use wearable sensing devices [7, 2, 13] that attach to the users' hand or body, but impose extra burden and inconvenience on users.

Recent radio-based gesture sensing technologies overcome the limitations of the above-mentioned approaches by using existing wireless infrastructure. WiSee [15] modifies IEEE 802.11 OFDM signal structure to extract Doppler patterns caused by movement of body or limbs, thus identifying certain gestures. Wi-Vi [1] uses virtual multi-antenna technology to identify signal fluctuations caused by body movement. The wide coverage and multipath reflections inherent to WiFi enable these solutions to work even under non-line-of-sight (NLOS) conditions. However, WiSee and WiVi can only recognize a selected set of coarse-grained gestures or whole-body movements, such as punching, kicking, and stepping forward/backward. In addition, they rely on sophisticated signal features that can be extracted using software radios, but are not readily provided by commercial off-the-shelf (COTS) IEEE 802.11 devices.

In this paper, we seek to advance wireless gesture sensing by answering the following question: is it possible to achieve finer-grained gesture sensing by simply using feature information that is readily available on COTS wireless devices? Specifically, we focus on wireless devices equipped with directional or beamforming antennas specified in standards such as the IEEE 802.11ad and IEEE 802.11n/ac. We target gestures that involve hand movement and no body movement, specifically, those in the standard American Sign Language (ASL). Our hypothesis is that by actively altering the directional wireless channel between the transmit and the receive antenna, these gestures create more significant changes in the signal feature such as Received Signal Strength (RSS) as compared to conventional omni-directional antennas.

To verify this hypothesis, we perform a set of baseline tests that require users to perform gestures between a pair of short-range directional WiFi transmitter and receiver. The test set involves application-specific, as well as randomly selected gestures from the ASL lexicon. Comprehensive experiments demonstrate that even with simple RSS based signal features, those fine-grained gestures exhibit clear patterns and can be identified with high probability. When used in conjunction with additional features such as targeted frequency subbands, temporal changes in phase differences in the selected frequency subbands, and RSS within each selected frequency

Ubicomp '14, September 13 - 17 2014, Seattle, WA, USA

http://dx.doi.org/10.1145/2632048.2632095

subband, we can further improve accuracy and granularity of gesture sensing.

Realizing the potential of a gesture recognition scheme in a real system involves a variety of design choices and challenges. The system must detect gesture onset and isolate the well-known fading effects that cause signal fluctuations. It should also be robust against unpredictable variations that may naturally occur in practical use cases. For example, user orientation relative to antennas may vary; the relative orientations of the transmit and receive antennas may not be perfectly aligned, and different users may perform the same gesture in slightly different ways.

We meet the above challenges through a practical and simple set of solutions. Our key idea is to match the feature information from users' gestures with a standard set of templates, while manipulating the matching algorithm to make it resilient to users' inconsistent behaviors. We also take advantage of the signal diversity coming from frequency-dependent fading to select the most reliable set of frequency bins for gesture sensing.

We implement the smart-antenna assisted gesture recognition framework on a 802.11-compatible radio platform that is equipped with directional antennas and can provide RSS and phase information at receiver side. We evaluate the framework in two field-test cases: on-wheelchair control and operations, and single-hand control of car entertainment system. For all the tests, we adopt a set of standard ASL containing around 25 gestures each, rather than a set of customized gestures that happen to favor our system. The set of ASL gestures include direction oriented gestures such as up, down; actions oriented such as start, stop; and nouns such as radio and television. Note that, these gestures require a fine-grained recognition system because they only consist of finger and/or hand movement. Our experiments demonstrate that a single instance of these gestures can be recognized with an accuracy of above 80% in most cases. Accuracy of recognition can be further increased by requiring an user to perform redundant gestures and/or by taking advantage of the context in which the gestures are performed. In addition, the performance is maintained when different users are involved in the test.

In summary, this paper makes the following contributions:

- We propose to leverage commonly available feature information from directional antennas for fine-grained gesture recognition, and provide a proof-of-concept evaluation.
- We design a set of practical algorithms that realize the directional antenna based gesture recognition in realistic wireless environments.
- We conduct comprehensive experiments to verify the effectiveness of the proposed approach in two fields of applications, using randomly selected ASL gestures and application-specific ASL gestures.

The rest of the paper is structured as follows. We start with a brief overview of relevant background information, and then present the detailed design of the proposed approach. Experimental evaluation of the proposed scheme is presented next. After a comparison to related research, a brief summary and future work is presented, and then we conclude this paper.

BACKGROUND

In this section, we provide background about the channel characteristics and antenna architectures in contemporary WiFi devices. These profiles are essential for enabling our gesture recognition system in practice.

A Primer on Wireless Channel Characteristics

Current WiFi devices commonly communicate through the orthogonal frequency division modulation (OFDM). An OFDM transmitter splits its spectrum band into multiple frequency subbands, called *subcarriers*, and sends the digital bits through these subcarriers in parallel. The digital bits are packetized and prepended with a *preamble*, which is used by the receiver to detect the starting point of each packet. With the preamble, the receiver can also estimate the channel distortion effects for each subcarrier separately. The channel distortion involves both magnitude attenuation and phase shift, which can be represented as a complex number, referred to as *channel state information* (CSI). It is necessary to estimate the CSI for each subcarrier separately, because practical radio environment tends to experience frequency-dependent channel distortion, due to multipath fading effects, i.e., signals with different wavelengths cancelling/strengthening each other in different ways [25].

WiFi device drivers usually expose the per-packet received signal strength (RSS) information to applications. Such RSS is essentially the total power of signals received across all subcarriers. Recent WiFi chipsets, such as Intel 5300 [5] and Atheros 9390 [21], further make per-subcarrier CSI available. Our gesture recognition system harnesses such rich channel information in modern WiFi devices to achieve high accuracy.

Directional Antenna

Emerging generations of WiFi devices are gradually incorporating smart antennas to improve communication capacity and reliability. 802.11n and 802.11ac compatible devices can be equipped with up to 4 and 8 antennas, respectively. Multiple transmit antennas can be digitally controlled to focus their beam power on a single receiver, or a selected set of receivers. The 802.11ad devices, running on the ultra-high frequency band of 60 GHz which enables miniature antennas, can assemble a even larger number of antennas on the same platform. These antennas can form a highly focused beam, as narrow as 2.81 degrees, towards the intended receiver [6].

Signals sent through such a narrow beam are highly directional, and resemble the properties of visible light, which is highly sensitive to blockage and multipath reflections. Practical 802.11ad transmitter and receivers tend to be placed in short-range, line-of-sight (LOS) setting to avoid such adverse effects. However, we observe that the sensitivity to such effects can be harnessed to improve the granularity of gesture recognition, when the gestures are performed in between the transmitter and receiver to intentionally create blockage and multipath reflections.

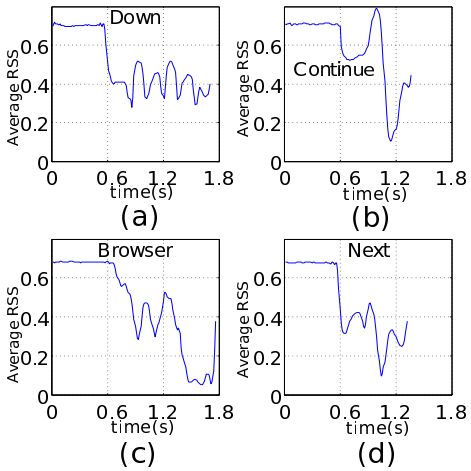

Figure 1. Average RSS corresponding to 4 randomly selected gestures.

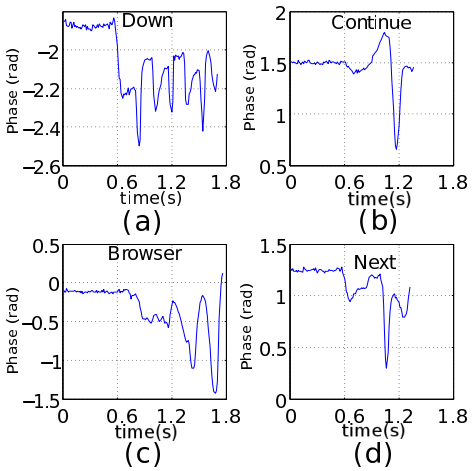

Figure 2. Phase variation corresponding to 4 randomly selected gestures.

As mentioned earlier, a major advantage of using the WiFi spectrum is that it does not have the potential to leak sensitive private information beyond the set of gestures performed. Furthermore, due to the inherent market of WiFi communication technologies and the use of commercial-off-the-shelf technology in our approach, the cost of our fine-grained gesture recognition system is expected to be low. The need for LOS deployment does, however, tend to limit the set of potential applications as compared to a non-LOS deployment.

A MEASUREMENT BASED FEASIBILITY STUDY

In this section, we motivate the proposed approach using illustrative experiments from one envisioned gesture-recognition application – the case of on-wheelchair control. A directional transmit antenna is mounted in a tripode on the left side of a user sitting on a chair, and a directional receive antenna is mounted in a tripode on the right side, emulating a wheelchair setup. Both antennas have a beamwidth of 30° and are mounted at a height of 80 cm. These two antenna are individually connected to a WARP v3 board running a typical IEEE 802.11g protocol stack[1]. WARP allows us to obtain detailed per-subcarrier CSI along with the total RSS across all subcarriers, in a similar way to COTS WiFi cards [5, 21]. During the experiments, the transmitter and receiver exchange random data packets using the IEEE 802.11g communication protocol. A person sitting in the chair performs 25 gestures taken from the American Sign Language (ASL). A typical ASL gesture comprises a unique hand/finger shape and movement of hands within a small region (several inches' distance). Our gesture set comprehensively covers commonly used ASL commands for controlling the wheelchair itself and ambient appliances.

In the case of the on-wheelchair control scenario we selected the following set of 25 gestures:

- Objects related: air conditioner, body, leg, fan, radio, phone, television, refrigerator, browser, window, washer, computer, kitchen;
- Direction related: left, right, up, down;
- Actions related: start, stop, next, previous, zoom, open, close, continue.

Figure 1(a)–(d) show the temporal variations in the total RSS of all subcarriers when four different gestures disturb the channel between the transmit and receive antenna. Similarly, Figure 2(a)–(d) show the temporal variations in the phase of the received signal on one of the selected subcarriers for the same gestures. It is visually clear that the temporal changes in the total RSS and phase have characteristics that distinguish the four gestures from each other. Therefore, total RSS and phase are possible features for recognizing a large number of gestures from ASL.

Unfortunately, the best distinguishing feature varies from gesture to gesture. This is illustrated in Figure 3. We evaluate the classification accuracy in a cross-validation experiment involving the aforementioned 25 gestures (x-axis). The y-axis plots the percentage of attempts where a gesture is correctly identified. For each gesture, we use three different features to run the cross-validation separately: average RSS, per-subcarrier RSS, and per-subcarrier phase. The same person performed the gesture for both training and testing data. Observe that, although the accuracy is good for all three features, the best feature for each gesture is different. In particular, for the 'Down' gesture the best feature is total RSS while for the 'Next' gesture, the best feature is RSS per subcarrier.

An additional note of importance regarding possible features is shown in Figure 4, where we examine the impact of gestures on the phase and RSS of each subcarrier separately. Figure 4(a)–(b) plot the classification accuracy for four gestures using the temporal variations in phase and RSS, respectively, on 5 different subcarriers. Note that, for different gestures, recognition accuracy varies across subcarriers, depending on if phase or RSS is used as features. For example, we can observe that the RSS of subcarrier 1 (SC-1) reaches much higher accuracy that the phase of SC-1 for recognizing the 'Next'

[1]The latest version of WARP driver [12] is fully compatible with 802.11g and can directly communicate with WiFi devices such as smartphones. The WARP boards were used only for the convenience of conducting the experiment and not because it offers any extra capability as compared to COTS WiFi hardware.

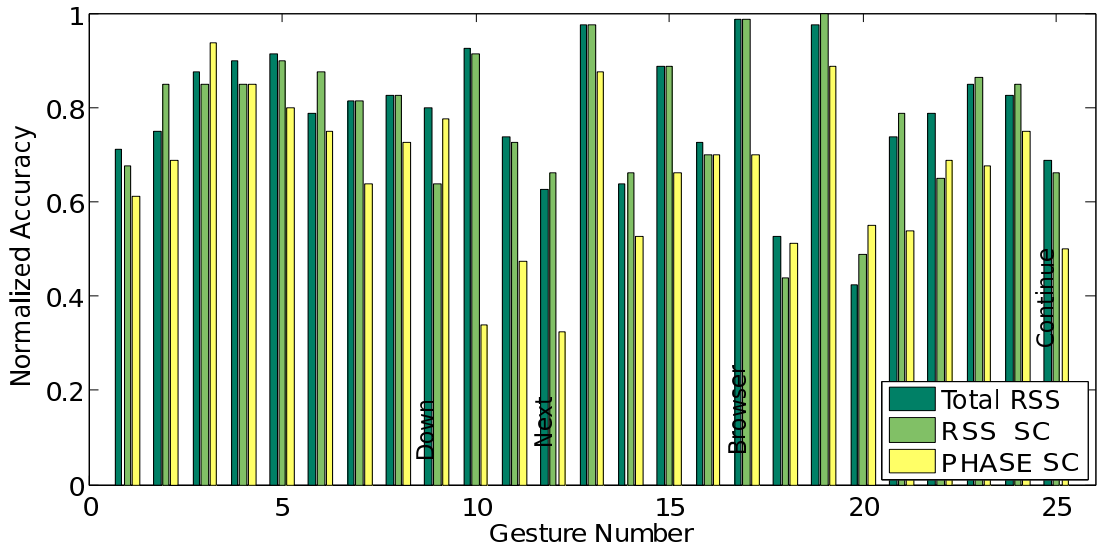

Figure 3. Accuracy of 25 Gestures considering 3 features for the classification: Average RSS, RSS of a specific subcarrier (SC), and phase of a specific subcarrier (SC).

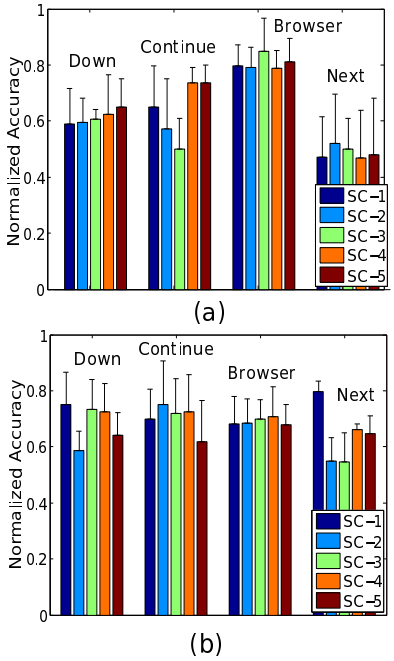

Figure 4. (a) Evaluation of the accuracy of different Phase Subcarriers Selection over 4 randomly selected Gestures. (b) Evaluation of the accuracy of different RSS per Subcarrier Selection over 4 randomly selected Gestures

gesture, whereas it is the other way around on SC-5 for the 'Continue' gesture. This shows that it is important to include selected subcarriers in the set of features used for classifying the gestures. Thus, our proposed approach selects the best of features for classifying gestures out of the collected feature data.

GESTURE SENSING ALGORITHM DESIGN

Above we have shown that it is feasible to visually distinguish different patterns when different gestures are performed. This intuition can be codified into signal features which can be extracted from either the average or total RSS (Figure 1) of all subcarriers, or RSS and phase per subcarrier (Figure 2) of the received signal. Therefore, a successful gesture recognition algorithm needs to effectively perform well in two main tasks:

- **Effective detection of the gestures**: This algorithm segments the received radio frequency (RF) signal and isolates the gesture feature from other signals.
- **Correct classification of the gestures**: The algorithm is able to distinguish between RF-based gesture signatures, by performing an appropriate pattern matching, and also it is resilient against gesture variations.

In order to accomplish theses tasks we develop a layered framework comprised of a gesture detection algorithm and gesture classification (pattern matching) algorithm, which we detail below.

Gesture Detection

The gesture detection algorithm plays an important role in the overall performance of the gestures identification system. It is important to detect the gestures first because in this way we can properly label the data we collect and train the system. Incorrect segmentation of the received signals can trigger incorrect classifications during the pattern matching. In addition, through gesture detection, we segment the signal samples we collected, so that only those segments corresponding to a gesture will be processed using the Nearest Neighbor pattern matching algorithm. This naturally reduces the computational cost. We remark that the WARP platform is used only for collecting the samples and host application for processing them. They are not imposing any constraint on our choice of the layered system architecture that isolates gesture detection from gesture recognition.

The segmentation algorithm identifies the starting and finishing points of a gesture pattern by looking at "silences" between gestures. These silence periods can correspond to a plateau response of the RF signal disturbed by a sudden change in the level of the signal, as a result of the hands'

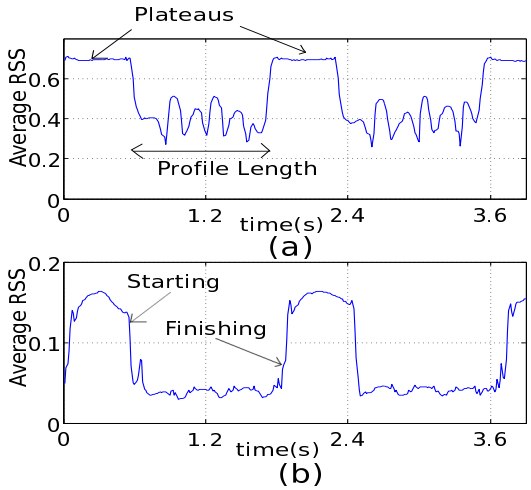

Figure 5. Example of gesture profiles in two different setups. In the case of (a), plateau zones help to identify the starting and finishing of a gesture. For the case of (b), we look at sharp level variation for identifying the point of interest.

initial blockage of the wireless link. Figure 5 shows two typical profiles. In particular, Figure 5(a) shows an example case where the identification of plateaus or silence zones between two consecutive gestures allows the profile extraction. Looking at Figure 1(a) and Figure 2(a), we can appreciate that plateau responses can be observed in the average RSS and in the RSS/phase per subcarrier. We can notice that the beginning of the gesture (after a flat response) are roughly synchronized among all the features, e.g., at around 0.6 second in the case of the "Down" gesture in both average RSS and per-subcarrier RSS. Therefore, for simplicity and computation time saving, we only leverage the average RSS in the segmentation algorithm.

For each gesture under training the detection of plateau periods is performed by first conditioning the time series of signals using a low pass filter (LPF). This allows the acquisition of a flatter signal during the silence period. When the signal does not change more that a threshold T for a t_s period of time, we identify either the starting or finishing point of a gesture. In order to distinguish between flat responses during silence periods and during gesture performing periods, we only consider those responses that exceed 60% of the average plateau levels of the signal RSS.

Figure 5 (b) shows a case where the identification of plateaus between two consecutive gestures does not help in the profile extraction. When silence periods are shorter than t_s, we are not able to identify the points of interest by using the logic just described. In these cases we look at sudden decays (when the gesture starts) and increases (when the gesture finishes) of the signal level. For identifying these points we use a combination of the first and second derivatives of the RSS time series. Values of the first derivative that exceed an empirical threshold T_d are deemed as candidates for starting or finishing points. For these candidates the joint use of the second derivative allows us to estimate if the RSS value corresponds to a maximum or minimum. Since gesture durations taken from the American Sign Language (ASL) have approximately similar durations, we can discharge many false alarm starting and finishing points, by just defining a minimum window of size equal to the average gesture duration which in our case corresponds to 92 time units or $1.8s$. This time is calculated from our experimental data.

Our actual system implementation runs the above two detection algorithms in parallel. A gesture will be segmented and its corresponding signals extracted if either algorithm detects it.

Gesture Training

Figure 6 illustrates the work flow of our gesture training algorithm. The inputs correspond to time series of average RSS per packet, and RSS/phase per subcarrier. In the case of the average RSS input the 'Gesture Segmentation' block finds the right starting and finishing points of a current gesture during the training process. Those points are translated into indexes that are taken as inputs for the 'Gesture Profile Generation' block. In the case of RSS and phase per-subcarrier input, every time series is processed by a 'Similarity Matrix' that selects the subcarriers of interest for the 'Gesture Profile Generator' block. This block generates all the gesture profiles that are stored in a Data Farm.

A typical gesture profile comprises 1 average-RSS feature vector, a number of (say C_1) per-subcarrier RSS feature vectors, and C_2 per-subcarrier phase feature vectors. Thus, the gesture is characterized using $(1+C_1+C_2)$ independent vectors of length τ, where τ corresponds to the length of the time series.

Gesture Pattern Matching

After acquiring the feature data and detecting the gesture segments, we need to compare the profile of the gestures under test against the profile of the reference gestures. This requires a method for evaluating the similarities between two gestures profiles P_i and P_j. One of the main challenges in the gesture pattern matching is to build an algorithm that is resilient to gesture variations. Even for the same person, he/she may perform the same gesture at very different speed and spacing, resulting in time-domain compression/stretch of the gesture pattern. An example of this situation is presented in Figure 5, where two average RSS time series are plotted. The gesture profiles in (a) and (b) correspond to exactly the same ASL sign. However, the lengths of these two expressions are quite different. Thus, regular Euclidean distance would not give the right metric for the classification. To overcome this limitation we use two techniques. The first approach is based on a cross-correlation method, and the second approach is based on Dynamic Time Warping (DTW) [10, 19] that is conventionally used in speech recognition.

Cross-correlation based method: For finding the level of similarity between two time series we use a normalized cross-correlation calculation with different lengths. The time series under testing correspond either to the average RSS, RSS per carrier or phase per carrier. Given two gesture time series P_i and P_j, the distance metric between them $\hat{D}_{(P_i,P_j)}(m)$ is calculated as:

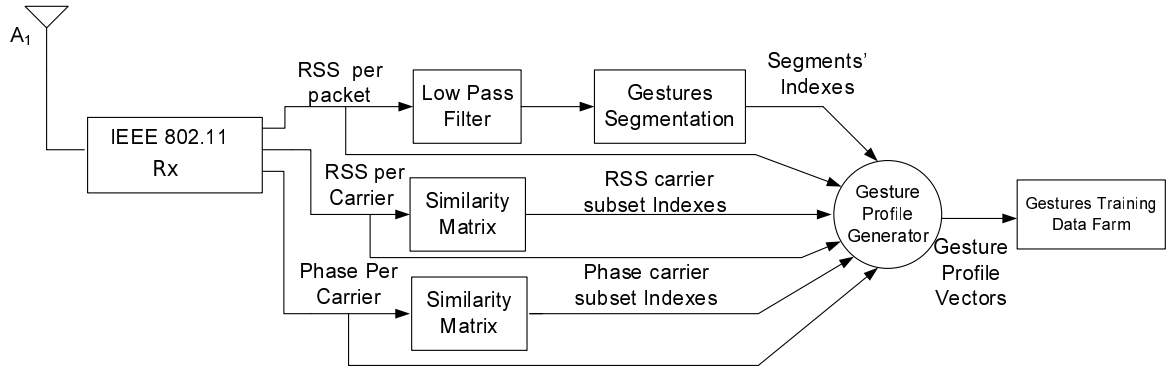

Figure 6. Gesture Training Algorithm

$$\hat{D}_{P_i,P_j}(m) = \begin{cases} \sum_{n=0}^{N-m-1} \frac{P_i(n+m)\times P_j^*(n)}{\sqrt{P_i(n+m)\times P_j(n)}} & \text{if m} \geq 0 \\ \hat{D}^*_{P_j,P_i}(-m) & \text{if m} < 0 \end{cases} \quad (1)$$

The value m in Eq. (1) accounts for the time lag that we consider for different time matching of the two series, and N length of the time series of the longer-duration gesture. The shorter time series is zero-padded to match the length of the longer-duration gesture.

Dynamic time warping (DTW) method: DTW can efficiently find the optimal alignment between two time series [19], even if they are stretched or compressed. In the same fashion as the cross-correlation based method, the time series under testing corresponds either to the average RSS, RSS per carrier or phase variation per carrier. Given two gesture time series P_i and P_j, and a Euclidean distance metric (2), DTW finds an alignment that matches each point in the first series to one or more points in the second series, such that the total distance of the matching summed over all point pairs is minimized. DTW finds the best alignment using standard dynamic programming [19].

$$D_{\alpha,\beta} = \|P_i(\alpha) - P_j(\beta)\| \quad (2)$$

Where α and β each represents an index from the time series P_i and P_j, respectively. DTW is equivalent to finding the α and β that minimizes the distance matrix (3):

$$\underset{O}{\text{minimize}} \quad \sum_{l=1,k=1}^{L} D_{\alpha_l,\beta_k} \quad (3)$$

where O represents a set of matrix elements in $D : O = \{\alpha_1.\beta_1\}, \{\alpha_2.\beta_1\}...\{\alpha_L.\beta_L\}$ and L corresponds to the length of the largest gesture profile.

Algorithm: Figure 7 shows the work flow of our gesture classification algorithm. Our baseline approach calculates the cross-correlation between the gesture under testing and all the reference gestures for classification, using several lags m. For simplicity the lags can be defined as $m = 1, 2..., 2N - 1$, where N corresponds to the length of time series of the longer-duration gesture. The best matching corresponds to the one with the maximum normalized cross-correlation value (1). Since we consider as features the average RSS, RSS per carrier and phase per carrier, the most reliable matching corresponds to the highest cross-correlation value for the same time series among these three feature profiles.

We define a level of confidence η that is the difference between the best and the second best cross-correlation value. The intuition behind this is to avoid miss-classification of gestures with very close RF profiles. If η is less than 10%, we use a computationally more expensive DTW algorithm, instead of cross-correlation, for the classification. In the case of DTW, we use the distance associated to the best DTW alignment between the two time series as metric in our classification algorithm. The best match between the test gestures and reference gestures corresponds to the one with the minimum DTW Distance.

Subcarrier selection: Harnessing frequency diversity

The OFDM based 802.11 communication protocol allows us to leverage the frequency diversity. Per-subcarrier phase and RSS profiles can help to obtain a more accurate classification than just considering the average RSS. However, the inclusion of all subcarriers (64 in the case of IEEE 802.11g) is computational expensive and not optimal because some profiles for different subcarriers are quite similar and do not provide extra useful information for classification purposes. This is particularly true when we consider the channel correlation between neighboring subcarriers that tend to have similar channel state information (CSI) [28]. Thus, to take advantage of the frequency diversity in gesture classification, a subcarrier selection mechanism is necessary. We design such a mechanism based on the evaluation of a $C \times C$ *similarity matrix*, where C represents the total number of subcarriers of the system. This means that we consider $C \times G$ similarity matrices, where G is total number of gestures. The similarity matrix is the confusion matrix when we do "leave-one-out" cross validation. It involves using a single subcarrier's features as the validation data, and the remaining subcarriers' as the data to compare with. When using this confusion matrix we assure that each subcarrier in the sample is used just once as the validation data. Each element (i, j) of the matrix is calculated as the cross-correlation value between the pair of subcarriers

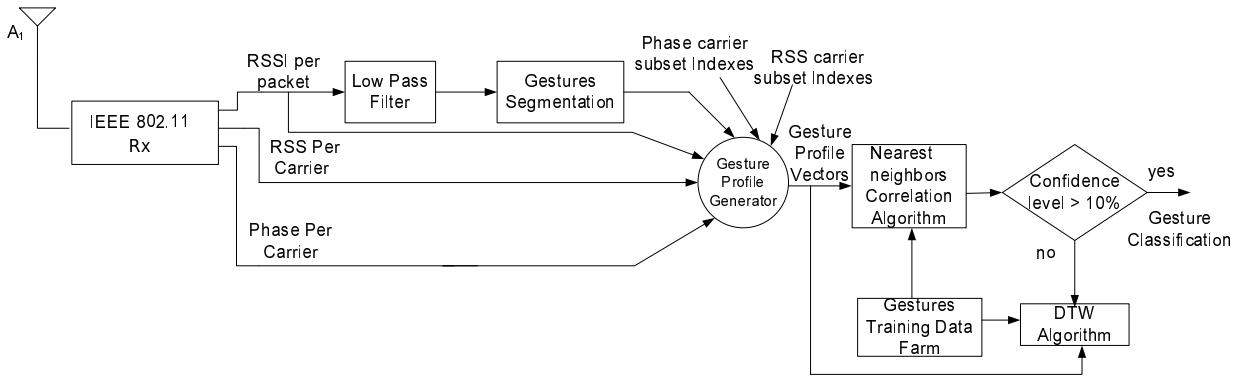

Figure 7. Work flow of the gesture classification algorithm

(i, j) for the same training set. From the similarity matrices we selected the 5 top subcarriers with the least average cross-correlation value among all gesture pairs. These subcarriers have potential to provide the most distinguishable profile of channel variation under different gestures.

EVALUATION

Experimental Setup

To verify the proposed approach we conducted comprehensive experiments in two different environment: (1) User sitting in a typical office chair, and (2) User sitting in the pilot seat of a conventional Car.

The first case emulates gesture-based electronic activation from a wheelchair. With this purpose in mind we setup a radio link using two directional antennas at each side of the chair, 72 cm apart and 80 cm above the floor, which correspond to a standard width and height of the armrests of a wheelchair. The chair is located in a typical office environment. Since this environment is characterized for having close-by walls, cubics, abundant other obstacles and people walking by, it represents a worst-case usage scenario with adverse multipath reflections that weaken the accuracy of gesture classification. We may expect less severe multipath reflection in outdoor deployment.

For the second scenario (Figure 8), a directional transmit antennae is mounted on the driver-side door of the automobile and a directional receive antennae is mounted on the front passenger-side door of the automobile. We tried to minimize as much as possible the visibility obstruction on both sides. For this experiment we use a typical Sedan car (Honda Civic), and the two antennas are separated by 132 cm. Since the interior of the car is enclosed and the propagation distance is large than the wheelchair scenario, we expect the adverse multipath that degrades the accuracy of the system to a larger extent.

Hardware

We set up a radio link using two WARP v3 boards, each connected to a RE14P directional patch antenna. The RE14P antenna bears a horizontal and vertical beam width of 30° [2]

[2]The form factor of the RE14P antenna is 19.5 cm × 19.5 cm and its usage is only limited to our prototype evaluation. Directional antennas in commercial WiFi products can be much smaller. For example, a 802.11ad 60 GHz highly-directional antennae array can be placed into a cm^2 area, owing to the milimiter-scale wavelength.

Figure 8. Experimental setup for gesture recognition inside a conventional Car.

We emphasize that the WARP driver we use is fully compatible [12] with 802.11g devices that can provide per-subcarrier channel profile information [5, 21] for gesture recognition. During the data collection process the transmitter and receiver communicate through the 802.11-compatible PHY layer running on WARP. RF signatures, including average RSS, RSS of each subcarrier and phase of subcarrier, are extracted from the WARP receiver and processed using a Matlab script running the gesture training/identification algorithms described above.

Data Collection and Evaluation

The goal of our experiment is to verify the accuracy of our fine-grained gesture recognition system. To this end, we profile each gesture using RF features, in particular average RSS, RSS and phase per selected subbands. We denote L as the number of RF samples we collect for each gesture. For a particular gesture i, we denote R_i as the average RSS vector, r_i^j and p_i^j as the RSS and phase, respectively, associated to the subcarrier j. This generates a combined vector $r_i^1, r_i^2, ..., r_i^C, p_i^1, p_i^2, ..., p_i^C \in \mathbb{R}^G$ when C subcarriers and G number of gestures are considered.

In total, for every profile gesture instance i there are $(2 \times C + 1) \times L$ values. We concatenate the signature vector of the i-th gesture and denote it as $s_i \in \mathbb{R}^{(2\times C+1)\times L}$. As a result, the whole data set for a given number of gestures G can be expressed as $S = \{s_i : i = 1, \cdots, G\}$.

We run an offline training phase to build the gesture templates and an online detection phase for identifying unknown ges-

tures. We emulate these two phases by separating the collected data into training and testing sets. We first partition the whole data set into five independent subsets of equal size $S_1, S_2, ..., S_5$. We use one subset as training set and the other four as testing set. For cross-validation purposes we repeat this process for each of the five sets. Thus, the resulting classification accuracy is the average among them. For each gesture in the testing set, we compare its RF signature vector s_i against the training set and return the nearest neighbor base in the metric defined in our gesture classification algorithm. Our total set size corresponds to 200 profiles per each gesture, or five subsets of 40 profiles per gesture.

Computation Time and Memory Requirements

Our gesture classification algorithm uses the Nearest Neighbor algorithm which has complexity $\mathcal{O}(Nd)$ with $N = 25$ and $d = 11(= 1 + 5 + 5)$ when we consider 25 gestures and 5 selected subcarriers. During the classification we need to perform cross-correlation calculations that in general can be hardware optimized. Also, computing the dynamic time warping operations bears $\mathcal{O}(N)$ complexity [19]. Therefore, we expect the computation can be handled in today's hardware platforms.

Memory requirements for our classification algorithm are low. Specifically, our system only requires storing 40 gestures profiles, with only 640KB of data per gesture.

Performance Evaluation

In this section, we evaluate the accuracy and stability of our gesture recognition framework under the aforementioned wheelchair and car settings.

Gesture recognition accuracy on a wheelchair. Figure 9 plots the confusion matrix when using our DTW-based algorithm to identify the 25 wheelchair-based gestures as listed in our feasibility study. A confusion matrix corresponds to a table layout that allows visualization of the recognition accuracy. In our case, each column of the matrix indexes a predicted gesture, while each row indexes an actual gesture performed by user. Element E_{ij} in the resulting matrix represents the probability that gesture i is identified as j. Ideally, the probability should be close to 1 along the diagonal. For example, in Figure 9 the (2,2) element of the matrix indicates that 92% of the time the predicted and actual gestures match, while the (3,2) element show that 3% of the cases the predicted gestures is the second (left) and the actual gesture corresponds to the third (body).

In Figure 9 we evaluate the baseline algorithm that randomly selects one subcarrier and uses its RSS profile as features. The results show that more than half of the gestures can be correctly identified with a probability larger than 0.81. Two out of all gestures experience large confusion, being correctly recognized with probability between 0.55 and 0.6.

However, with subcarrier selection enabled, the accuracy can be significantly improved, as shown in Figure 10. Among the 25 gestures, 22 can be correctly identified with probability larger than 0.9, and 11 are recognized with no errors. We also evaluate an evenly spaced carrier selection strategy, where we consider 5 non-null carriers spaced by 10 (carriers 10, 20, 30, 40, 50). In this case we obtain an average of 77% of accuracy among the 25 gestures, which is much lower than the average accuracy when using subcarrier selection strategy (92%). Therefore, the results clearly demonstrate the effectiveness of our approach.

Accuracy of gesture recognition in a car. We proceed to evaluate the in-vehicle scenario with 14 gestures. Figure 11 plots the resulting confusion matrix. We can see that half of the gestures achieves a recognition accuracy of larger than or equal to 0.89. Two gestures experience low accuracy, with detection probability of 0.62 and 0.65, respectively. Compared with the wheelchair scenario, such relatively lower accuracy is primarily due to larger separation between the transmit and receive antennas (72 cm vs. 132 cm), which leads to a larger beamwidth, and hence more severe multipath effects. Chambering the multipath reflections inside the small space induces more variations when the user performs the gestures.

One may wonder if adding more training data improves accuracy. Figure 12 shows the result when we increase the training set size up to 4 times. We can see that only a few gestures enjoy slight increase of accuracy, others' recognition accuracy may even drop. The Nearest Neighbor pattern matching algorithm that we use is inherently sensitive to noise, regardless of the training set size. Adding new training data does not necessarily improve accuracy. In fact, we risk adding more anomaly/noise points which reduces classification accuracy. This is why, as we increase the training set, the accuracy does not necessarily improve.

Evaluation of robustness. We now evaluate whether our gesture recognition approach is ready for spontaneous use case with imperfectly controlled settings. In particular, we aim to answer the following questions: (i) How sensitive is the gesture recognition to misalignment between the transmit and receive antennas? (ii) How will the accuracy be affected by the antennas' separation? (iii) Does the system perform consistently across different user?

Figure 13 shows the experimental results under such variations. The experiments are conducted under the wheelchair setting. Here "Training" represents the average accuracy across all 25 gestures and error bars the variance. This is derived from the confusion matrix in Figure 10) and used as a benchmark for robustness test. To evaluate the impact of antenna misalignment, we intentionally offset the antenna angles by ± 5 degrees, but still use the original training data to recognize the gestures under testing. The results show that the mean accuracy drops from 92% to 67%. Antenna misalignment causes the RSS pattern to deviate from the original one, thus disturbing the classification algorithm. In addition, some "leakage" signals may arrive at the receive antenna after significant reflections, which causes variation of RSS across subcarriers. In this case the accuracy can be further improved by using redundant gestures and/or by taking advantage of the context in which the gestures are performed.

Figure 13 also shows a reduction of accuracy when we intentionally vary the antenna distance by ± 10 cm. Under different separation, the overlapping of beams between the trans-

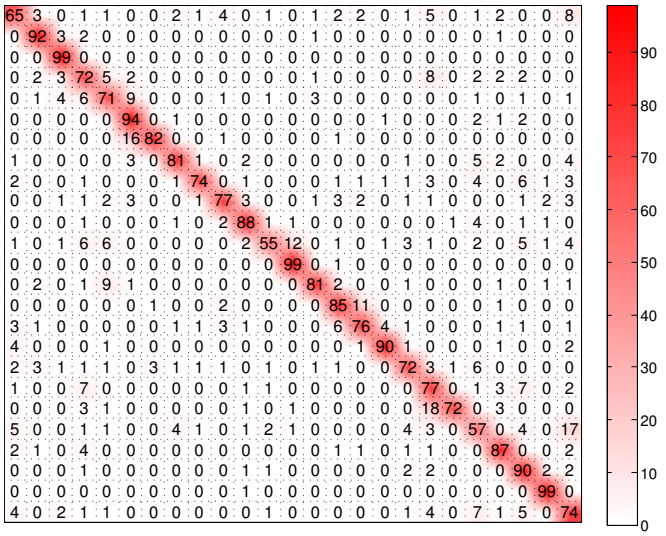

Figure 9. Confusion matrix for 25 gestures in wheelchair deployment when using DTW without subcarrier selection algorithm.

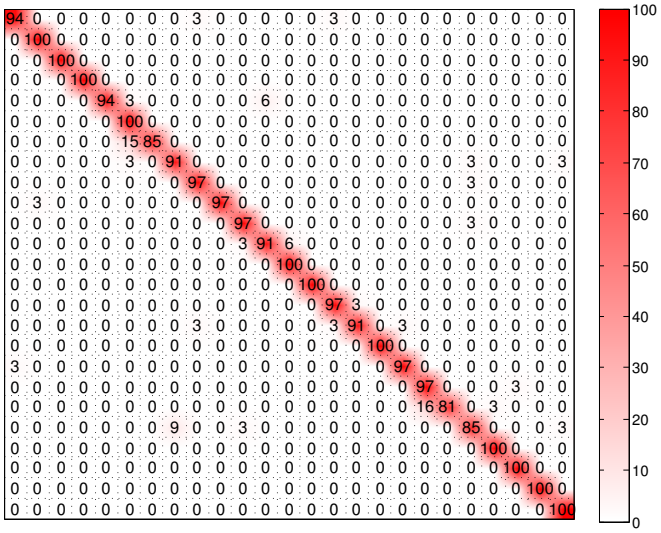

Figure 10. Confusion matrix for 25 gestures in wheelchair deployment when using DTW and subcarrier selection algorithm.

Figure 11. Confusion Matrix for 14 Gestures in Car Deployment when using the Greedy Classification Algorithm.

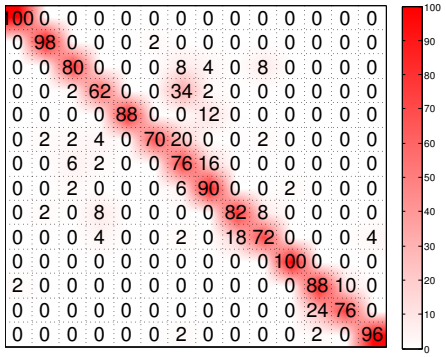

Figure 12. Confusion Matrix for the same 14 Gestures when using a 4 times longer Training Set.

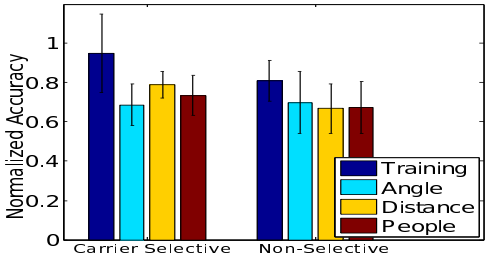

Figure 13. Robustness Study: Average accuracy for different Angles, Persons performing the gestures, and distance between antennas for 1. Subcarrier selective and 2. non-selective Case.

mit and receive antennas deviates from the original one, leading to different RSS patterns under each gesture. As a result, when we keep using the original training set, average accuracy can drop by around 16%.

We also request two additional users to test our system, by instructing them to follow the gestures of an experienced user. Figure 13 shows that these new users tend to experience 15% to 20% lower accuracy, when we attempt to match there gesture features with the experienced user's.

In summary, the above experiments reveal a tradeoff between gesture recognition accuracy and robustness. Although the variation of channel RSS/phase can be harnessed to recognize fine-grained gestures, it is sensitive to changes in the antenna placement and differences of gestures across users.

Notably, across all the experiments, our approach that combines subcarrier selection with feature classification achieves higher accuracy than that without selection.

RELATED WORK

The application space of gesture sensing is diverse, covering not only everyday scenarios, but also medical systems and assistive technologies, crisis management and disaster relief, etc. In this section, we will discuss existing approaches to gesture recognition and their application domains.

Vision based gesture recognition. Decades of research has been devoted in computer vision technologies that distinguish video-recorded gestures. Wahs et al. [26] presented a comprehensive survey of this line of research. Early in the 1990's, Rehg et al. [16] used stereo cameras to capture hand images, which are simplified into regular geometrical shapes as gesture templates. Using only a coarse description of hand shape

and a Hidden Markov Model (HMM), Starner et al. [23] are able to recognize 42 American Sign Language gestures with 99% accuracy. More recent work uses depth cameras to enable in-air 3D human computer interactions [18]. The highly populated Kinect device is a state-of-the-art example application. Vision based gesture sensing technologies have been applied in vehicles [14] to control the infotainment system and reduce driver distraction. Despite the vast literature, vision based approaches face some inherent challenges, e.g., coping with dynamic backgrounds, low-light or sunlight interference, high computational cost (long response time).

Sensor based gesture recognition. Wearable or near-body sensing devices can provide a rich set of signal features for gesture recognition. For example, magnetic sensors can extract gesture patterns from a moving hand attached with metal objects [8] [7]. Motion sensors (accelerometer and gyroscope) can identify fine-grained gestures, such as virtual characters made by moving a smartphone in the air [2, 13]. Observing that human body can be abstracted as a capacitor that disturbs ambient electronic signals, researchers exploited electronic field sensors [22, 3] for whole-body movement identification. These technologies have potential to relieve users from wearing special devices, but they are sensitive to environment and human dynamics (e.g., ambient electronic devices, thickenss of clothing, and contact with other persons) [14]. Recently, Leap Motion Inc. [11] launched a portable device that emits multiple channels of infrared signals and extracts patterns from the reflection to identify hand gestures and postures. Compared with such technologies, wireless gesture sensing has a low-cost and is less intrusive — it only needs support from readily available network infrastructures.

Audio and radio based gesture sensing. Audio signals generated by mobile devices may be reflected in different ways when human posture changes. The resulting pattern can be leveraged to sense the human activities [24, 20]. Similarly, when multiple radio links coexist, each can detect a certain signal variation when users present in between, and the variation patterns can be mapped to presence/movement information, a technology called radio tomography [29, 27]. An alternative approach [4] extracts the Doppler features from sound waves reflected by human gestures relevant to interaction with computers. WiSee extended this approach to WiFi [15] to enable long-range detection of 9 gestures involving whole-body or limbs. WiVi [1] further enables cross-wall whole-body movement sensing by migrating inverse synthetic aperture radar (ISAR) technology to WiFi devices. These approaches involve sophisticated customized signal processing algorithms that cannot be readily deployed in current wireless platforms. In addition, since they only rely on the binary change of Doppler frequency (negative or positive), only a limited set of coarse-grained gestures can be identified.

Summary: A feature comparison of gesture recognition systems. Table 1 compares our approach with state-of-the-art radio and vision based gesture recognition technologies. Similar to commercial vision based approaches such as Kinect and Leap Motion, our system realizes hand-level recognition granularity. The computational complexity (run time) of our system is lower than 2-D and 3-D video image processing used by Kinect. As an RF based system, it can work under any lighting conditions, whereas systems that work with infrared technology such as Kinect and Leap Motion may not work well in outdoor and bright environment [17, 9]. The detection distance of our system is larger than Leap Motion and similar to Kinect. Alternative RF-based approaches such as WiSee rely on Doppler effects which can only distinguish coarse-grained body/limb movement. Finally, since our system is based on IEEE 802.11 standard, the implementation does not require dedicated Hardware or modification of existing hardware.

	Our	WiSee [15]	Kinect	Leap Motion [11]
Granularity	Fine-Grained	Coarse-Grained	Fine-Grained	Fine-Grained
Run Time	Low	Low	High	Low
Lighting	Any	Any	Indoor Lighting	Indoor Lighting
Detection Distance	$\approx$ 6 ft	Multi Rooms	$\approx$ 10 ft	$\approx$ 1 ft
Cost	Low	Medium - High	High	Low
Dedicated Hw	No	yes	yes	yes

Table 1. A comparison of recent gesture recognition technologies.

CONCLUSION

In this paper, we have provided experimental evidence to demonstrate the feasibility of fine-grained gesture recognition using directional antennas. With the experimental insights, we develop a practical framework that extracts channel profile from commodity directional WiFi radios as signatures to distinguish gestures. In particular, we explore the frequency-selectivity of channel magnitude and phase to select the most distinguishing features to feed the pattern matching framework. We further design a dynamic time warping based algorithm to improve resilience of the gesture matching to variations of the user behavior. Our experimental evaluation shows that this approach can achieve an average of 92% recognition accuracy in a wheelchair usage case with 25 gestures from the American Sign Language, and 84% accuracy in a vehicle with 14 gestures. The results verify the potential of using unmodified WiFi devices to recognize fine-grained gestures.

Throughout our experiments, we have used a pair of directional antennas with a fixed beamwidth of 30 degrees. The emerging 60 GHz 802.11ad devices bear a much narrower beam and can further improve the accuracy and granularity of gesture recognition. In addition, the availability of multiple antennas in 802.11n/ac devices provides an additional dimension of diversity, which may enrich the signature space and enhance gesture recognition. We plan to explore such possibilities in our future work.

REFERENCES

1. Adib, F., and Katabi, D. See Through Walls with WiFi! In *Proc. of ACM SIGCOMM* (2013).

2. Agrawal, S., Constandache, I., Gaonkar, S., Roy Choudhury, R., Caves, K., and DeRuyter, F. Using Mobile Phones to Write in Air. In *Proc. of ACM MobiSys* (2011).
3. Cohn, G., Morris, D., Patel, S., and Tan, D. Humantenna: Using the Body As an Antenna for Real-time Whole-body Interaction. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2012).
4. Gupta, S., Morris, D., Patel, S., and Tan, D. SoundWave: Using the Doppler Effect to Sense Gestures. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2012).
5. Halperin, D., Hu, W., Sheth, A., and Wetherall, D. Predictable 802.11 Packet Delivery from Wireless Channel Measurements. In *Proc. of ACM SIGCOMM* (2010).
6. IEEE Standard. 802.11™: Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band, 2012.
7. Ketabdar, H., Moghadam, P., Naderi, B., and Roshandel, M. Magnetic Signatures in Air for Mobile Devices. In *Proc. of International Conference on Human-computer Interaction with Mobile Devices and Services Companion (Mobile HCI)* (2012).
8. Ketabdar, H., Roshandel, M., and Yüksel, K. A. Towards Using Embedded Magnetic Field Sensor for Around Mobile Device 3D Interaction. In *Proc. of International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)* (2010).
9. Kirk McElhearn. Leap Motion Controller Fails in Normal Conditions, http://www.mcelhearn.com/not-a-review-leap-motion-controller-fails-in-normal-conditions/, 2014.
10. Kruskal, J. B., and Liberman, M. The symmetric time-warping problem: from continuous to discrete. In *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, D. Sankoff and J. B. Kruskal, Eds. CSLI Publications, 1999.
11. Leap Motion, Inc. Leap Motion: Mac & PC Gesture Controller for Game, Design and More. `https://www.leapmotion.com/`, 2013.
12. Mango Communications. WARP 802.11 Reference Design. `http://warpproject.org/trac/wiki/802.11`, 2013.
13. Park, T., Lee, J., Hwang, I., Yoo, C., Nachman, L., and Song, J. E-Gesture: A Collaborative Architecture for Energy-efficient Gesture Recognition with Hand-worn Sensor and Mobile Devices. In *Proc. of ACM SenSys* (2011).
14. Pickering, C. A., Burnham, K. J., and Richardson, M. J. A Research Study of Hand Gesture Recognition Technologies and Applications for Human Vehicle Interaction. In *Institution of Engineering and Technology Conference on Automotive Electronics* (2007).
15. Pu, Q., Gupta, S., Gollakota, S., and Patel, S. Whole-home Gesture Recognition Using Wireless Signals. In *Proc. of ACM MobiCom* (2013).
16. Rehg, J., and Knade, T. Visual Tracking of High DOF Articulated Structures: an Application to Human Hand Tracking. In *European Conference on Computer Vision* (1994).
17. RoboRealm. Microsoft Kinect, http://www.roborealm.com/help/Microsoft_Kinect.php, 2013.
18. Saba, E., Larson, E., and Patel, S. Dante Vision: In-Air and Touch Gesture Sensing for Natural Surface Interaction With Combined Depth and Thermal Cameras. In *IEEE International Conference on Emerging Signal Processing Applications (ESPA)* (2012).
19. Salvador, S., and Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal. 11*, 5 (2007).
20. Scholz, M., Riedel, T., Hock, M., and Beigl, M. Device-free and Device-bound Activity Recognition Using Radio Signal Strength. In *Proc. of the Augmented Human International Conference* (2013).
21. Sen, S., Lee, J., Kim, K.-H., and Congdon, P. Avoiding Multipath to Revive Inbuilding WiFi Localization. In *Proc. of ACM MobiSys* (2013).
22. Smith, J. R. Field Mice: Extracting Hand Geometry from Electric Field Measurements. *IBM Systems Journal*, 3/4.
23. Starner, T., and Pentland, A. Real-time American Sign Language Recognition from Video Using Hidden Markov Models. In *Proc. of International Symposium on Computer Vision* (1995).
24. Tarzia, S. P., Dick, R. P., Dinda, P. A., and Memik, G. Sonar-based Measurement of User Presence and Attention. In *Proc. of ACM UbiComp* (2009).
25. Tse, D., and Viswanath, P. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
26. Wachs, J. P., Kölsch, M., Stern, H., and Edan, Y. Vision-based Hand-gesture Applications. *Communications of the ACM 54*, 2 (2011).
27. Wilson, J., and Patwari, N. Radio Tomographic Imaging with Wireless Networks. *IEEE Transactions on Mobile Computing 9*, 5 (2010).
28. Xie, X., Zhang, X., and Sundaresan, K. Adaptive feedback compression for mimo networks. In *Proc. of ACM MobiCom* (2013).
29. Zhao, Y., Patwari, N., Phillips, J. M., and Venkatasubramanian, S. Radio Tomographic Imaging and Tracking of Stationary and Moving People via Kernel Distance. In *Proc. of ACM/IEEE IPSN* (2013).