Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi

Yue Zheng[1], Yi Zhang[1], Kun Qian[1], Guidong Zhang[1], Yunhao Liu[1,2], Chenshu Wu[3], Zheng Yang[1*]
[1]Tsinghua University, China
[2]Michigan State University, USA
[3]University of Maryland, College Park, USA
{cczhengy,zhangyithss,qiank10,zhanggd18,yunhaoliu,wucs32,hmilyyz}@gmail.com

ABSTRACT

Wi-Fi based sensing systems, although sound as being deployed almost everywhere there is Wi-Fi, are still practically difficult to be used without explicit adaptation efforts to new data domains. Various pioneering approaches have been proposed to resolve this contradiction by either translating features between domains or generating domain-independent features at a higher learning level. Still, extra training efforts are necessary in either data collection or model re-training when new data domains appear, limiting their practical usability. To advance cross-domain sensing and achieve fully zero-effort sensing, a domain-independent feature at the lower signal level acts as a key enabler. In this paper, we propose Widar3.0, a Wi-Fi based zero-effort cross-domain gesture recognition system. The key insight of Widar3.0 is to derive and estimate velocity profiles of gestures at the lower signal level, which represent unique kinetic characteristics of gestures and are irrespective of domains. On this basis, we develop a one-fits-all model that requires only one-time training but can adapt to different data domains. We implement this design and conduct comprehensive experiments. The evaluation results show that without re-training and across various domain factors (i.e. environments, locations and orientations of persons), Widar3.0 achieves 92.7% in-domain recognition accuracy and 82.6%-92.4% cross-domain recognition accuracy, outperforming the state-of-the-art solutions. To the best of our knowledge, Widar3.0 is the first zero-effort cross-domain gesture recognition work via Wi-Fi, a fundamental step towards ubiquitous sensing.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Gesture Recognition; Channel State Information; COTS Wi-Fi

ACM Reference Format:
Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *The 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '19), June 17–21, 2019, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3307334.3326081

*Yue Zheng and Yi Zhang are co-first authors. Zheng Yang is the corresponding author.

MobiSys '19, June 17–21, 2019, Seoul, Republic of Korea

ACM ISBN 978-1-4503-6661-8/19/06…$15.00
https://doi.org/10.1145/3307334.3326081

1 INTRODUCTION

Human gesture recognition is the core enabler for a wide range of applications such as smart home, security surveillance and virtual reality. Traditional approaches use cameras [16, 24, 42], wearable devices and phones [8, 17, 36] or sonar [22, 29, 48] as the sensing module. While promising, these approaches pose inconvenience due to their respective drawbacks including leakage of privacy, requirement of on-body sensors and limit of sensing range. The need for secure, device-free and ubiquitous gesture recognition interface has triggered extensive research on sensing solutions based on commodity Wi-Fi. Pioneer attempts such as E-eyes [45], CARM [44], WiGest [1] and WIMU [38] have been proposed. In principle, early wireless sensing solutions extract either statistical features (e.g., histograms of signal amplitudes [45]) or physical features (e.g., power profiles of Doppler frequency shifts [44]) from Wi-Fi signals and map them to human gestures. However, these primitive signal features usually carry adverse environment information unrelated to gestures. Specifically, due to lack of spatial resolution, wireless signals, and their features as well, are highly specific to *environment* where the gesture is performed, and the *location and orientation* of the performer, as Figure 1 shows. For brevity, we unitedly term these factors irrelevant to gestures as *domain*. As a result, the classifiers trained with primitive signal features in one domain usually undergo drastically drop in accuracy with another domain.

Recent innovations in gesture recognition with Wi-Fi have explored cross-domain generalization ability of recognition models. For example, recent works [20, 50] borrow the ideas from machine learning, such as transfer learning and adversarial learning, and apply advanced learning methodologies to improve cross-domain recognition performance. Another solution, WiAG [39], derives a translation function to generate signal features of the target domain for model re-training. While to some extent achieving cross-domain recognition, all existing works require extra training efforts in either data collection or model re-training at each time a new target domain is added into the recognition model. Even worse, correlated with continuous location and orientation of a person, Wi-Fi signals have infinite number of domains, making cross-domain training approaches practically prohibitive.

A more promising but challenging solution is a *"one-fits-all"* model that is able to *train once, use anywhere*. Such ideal model, trained in one domain, can be directly used in new domains without extra efforts, such as data collection, generation, or re-training. Different from all existing approaches, our key idea is to move generalization ability downwardly at the lower signal level, rather than

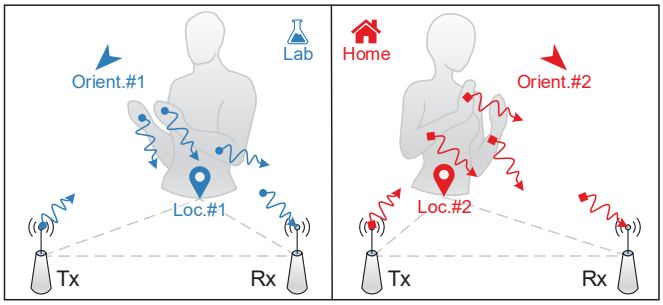


Figure 1: Cross-domain gesture recognition, where persons may be at different locations and orientations relative to Wi-Fi links, and environments (e.g., lab, home, etc.). In this example, one male and one female are performing clapping gestures in two domains.

the upper model level. Specifically, we extract domain-independent features reflecting only gesture itself from raw domain-dependent signals. On this basis, we aim to build an explainable cross-domain recognition model that can be applied in new scenarios with zero effort and high accuracy.

However, we face three major technical challenges to achieve a *one-fits-all* model. First, previously used signal features (e.g., amplitude, phase, Doppler Frequency Shift(DFS)), as well as their statistics (e.g., max, min, mean, distribution parameter), are absolutely domain-dependent, meaning that their values vary with different locations, orientations and environments even for the same gesture. Second, it is difficult, for radio signals from only several links, to describe human gestures and actions. For example, kinetic profile of a single gesture still has hundreds of variables, posing the estimation of kinetic profile as a highly under-determined problem. Third, cross-domain generalization often requires sophisticated learning models (e.g., deeper networks, a larger number of parameters, a more complex network structure and more complicated loss functions), which slow down or even obstruct training, over-consume training data, and make the model less explainable.

To overcome these challenges, we propose Widar3.0, a Wi-Fi based gesture recognition system. Widar3.0 uses channel state information (CSI) portrayed by COTS Wi-Fi devices. Our prior efforts, Widar [32] and Widar2.0 [33] track coarse human motion status, e.g., location and velocity, by regarding a person as a single point. Widar3.0, however, aims at recognizing complex gestures that involve multiple body parts. The key component of Widar3.0 is our novel theoretically domain-independent feature *body-coordinate velocity profile* (BVP) that describes power distribution over different velocities, at which body parts involved in the gesture movements. Our observation is that each type of gestures has its unique velocity profile in the body coordinate system (e.g., the coordinates where the orientation of the person is the positive x axis) no matter in which domain is the gesture performed. To estimate BVP, we approximate BVP from several prominent velocity components and further employ compressive sensing techniques to derive accurate estimates. On this basis, we devise a learning model to capture spatial-temporal characteristics of gestures and finally classify gestures. Through downward movement of model generalization techniques closer to the raw signals, Widar3.0 enables zero-effort cross-domain human gesture recognition with many expected properties simultaneously, including high and reliable accuracy, strong generalization ability, explainable features, reduced amounts of training data. We implement Widar3.0 on COTS Wi-Fi devices and conduct extensive field experiments (16 users, 15 gestures, 15 locations and 5 orientations in 3 environments, and comparisons with three state-of-the-art approaches). Especially, the results demonstrate that Widar3.0 significantly improves the accuracy of gesture recognition to 92.4% in cross-environment cases, while the recognition accuracy with raw CSI and DFS profiles are 40.2% and 77.8% only. Across different types of domain factors including user's location, orientation, environment and user diversity, Widar3.0 achieves average accuracy of 89.7%, 82.6%, 92.4% and 88.9%, respectively.

In a nutshell, our core contributions are three-fold. First, we present a novel domain-independent feature that captures body-coordinate velocity profiles of human gestures at the lower signal level. BVP is theoretically irrespective of any domain information in raw Wi-Fi signals, and thus acts as a unique indicator for human gestures. Second, we develop a *one-fits-all* model on the basis of domain-independent BVP and a learning method that fully exploits spatial-temporal characteristics of BVP. The model enables cross-domain gesture recognition without any extra effort of data collection or model re-training. Third, though trained only once, Widar3.0 achieves on average 89.7%, 82.6%, and 92.4% recognition accuracy across locations, orientations, and environments, respectively, which outperform the state-of-the-art solutions that require re-training in new target domains. Such consistently high performance demonstrates its strong ability of cross-domain generalization. To the best of our knowledge, Widar3.0 is the first zero-effort cross-domain gesture recognition via Wi-Fi, a fundamental step towards ubiquitous sensing.

2 MOTIVATION

Widar3.0 addresses the problem of cross-domain gesture recognition with Wi-Fi signals. Due to the lack of spatial resolution, wireless signals are highly formatted by domain characteristics. Either or not to some extent enabling cross-domain sensing, existing wireless sensing solutions have significant drawbacks in their feature usage. The three main types of features are listed as follows:

Primitive features without cross-domain capability. Most state-of-the-art activity recognition works extract primitive statistical (e.g., power distribution, waveform) or physical features (e.g., DFS, AoA, ToF) from CSI [46]. However, due to different locations and orientations of the person and multipath environments, features of the same gesture may vary significantly and fail to serve successful recognition. As a brief example, a person is asked to push his right hand multiple times, yet with two orientations relative to the wireless link. The spectrograms are calculated as in [44], and dominant DFS caused by the movement of the hand is extracted. As shown in Figure 2, while dominant DFS series of gestures with the same domain form compact clusters, they differ greatly in trends and amplitudes between two domains, and thus fail to indicate the same gesture.

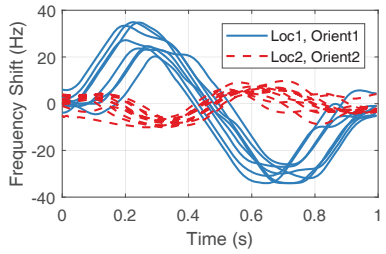


Figure 2: Dominant DFS of gesture differs with person orientations and locations.

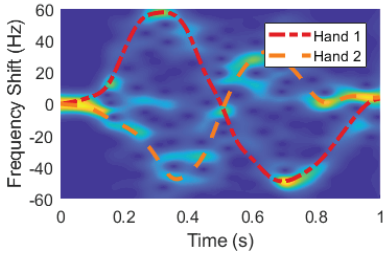


Figure 3: Complex gestures cause multiple DFS components.

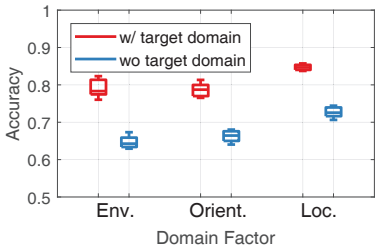


Figure 4: Accuracy of adversarial learning drops without target domain data.

Cross-domain motion features for coarse tracking. Device-free tracking approaches [26, 33] build quantitative relation between physical features of signal and the motion status of the person, and enable location and velocity measurement across environments. However, these works regard a person as single point, which is infeasible for recognizing complex gestures that involve multiple body parts. Figure 3 illustrates the spectrogram of a simple hand clap, which contains two major DFS components caused by two hands and a few secondary components.

Latent features from cross-domain learning methods. Cross-domain learning methods such as transfer learning [50] and adversarial learning [20] latently generate features of data samples in the target domain, either by translating samples from the source domain, or learning domain-independent features. However, these works require extra efforts of collecting data samples from the target domain and retraining the classifier each time new target domains are added. As an example, we evaluate the performance of an adversarial learning based model, EI [20] over different domain factors (e.g., environment, location and orientation of the person). Specifically, the classifier is trained with and without data samples in every type of target domains. As shown in Figure 4, the system accuracy obviously drops without the knowledge of the target domains, demonstrating the need of extra data collection and training efforts in these learning methodologies.

Lessons learned. The deficiency of existing cross-domain learning solutions asks for a new type of domain-independent feature. Should it be achieved, a *one-fits-all* model could be built upon it to save much data collection and training efforts. Widar3.0 is designed to develop and exploit body-coordinate velocity profile (BVP) to address the issue.

3 OVERVIEW OF WIDAR3.0

Widar3.0 is a cross-domain gesture recognition system using off-the-shelf Wi-Fi devices. As shown in Figure 5, multiple wireless links are deployed around the monitoring area. Wireless signals, as distorted by the user in the monitoring area, are acquired at receivers and their CSI measurements are logged and preprocessed to remove amplitude noises and phase offsets.

The major parts of Widar3.0 are two modules, the *BVP generation module* and the *gesture recognition module*.

Upon receiving sanitized CSI series, Widar3.0 divides CSI series into small segments, and generates BVP for each CSI segment via the BVP generation module. Widar3.0 first prepares three intermediate results: DFS profiles, the orientation and location information of the person. DFS profiles are estimated by applying time-frequency analysis to CSI series. The orientation and location information of the person is calculated via motion tracking approaches. Thereafter, Widar3.0 applies the proposed compressed-sensing-based optimization approach to estimate BVP of each CSI segment. The BVP series is then output for following gesture recognition.

The gesture recognition module implements a deep learning neural network (DNN) for gesture recognition. With the BVP series as input, Widar3.0 performs normalization on each BVP and across the whole series, in order to remove the irrelevant variations of instances and persons. Afterwards, the normalized BVP series is input into a spatial-temporal DNN, which has two main functions. First, the DNN extracts high-level spatial features within each BVP using convolutional layers. Then, recurrent layers are adopted to perform temporal modeling of inter-characteristics between BVPs. Finally, the output of the DNN is used to indicate the type of the gesture performed by the user. In principle, Widar3.0 achieves zero-effort cross-domain gesture recognition, which requires only one-time training of the DNN network, but can be directly adapted to as many as new domains.

4 BODY-COORDINATE VELOCITY PROFILE

Intuitively, human activities have unique velocity distributions across all body parts involved, which can be used as activity indicators. Among all parameters (i.e. ToF, AoA, DFS and attenuation) of the signal reflected by the person, DFS embodies most information of velocity distribution. Unfortunately, DFS is also highly correlated with the location and orientation of the person, circumventing direct cross-domain activity recognition with DFS profiles.

In this section, we tempt to derive distribution of signal power over velocity components in the body coordinate system, i.e. BVP, which uniquely indicates the type of activities. Preliminary of the CSI model is first introduced (§ 4.1), followed by the formulation and calculation of BVP (§ 4.2 and § 4.3). Finally, prerequisites for calculating BVP are given (§ 4.4).

4.1 Doppler Representation of CSI

CSI portrayed by off-the-shelf Wi-Fi devices describes multipath effects in the indoor environment at arrival time t of packets and

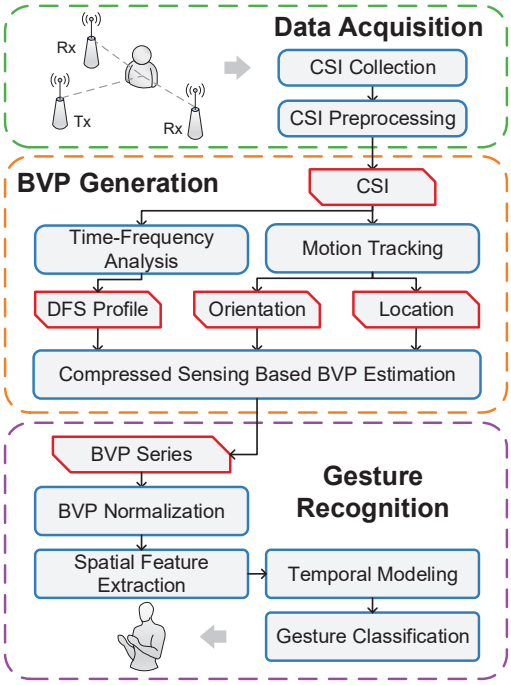


Figure 5: System overview.

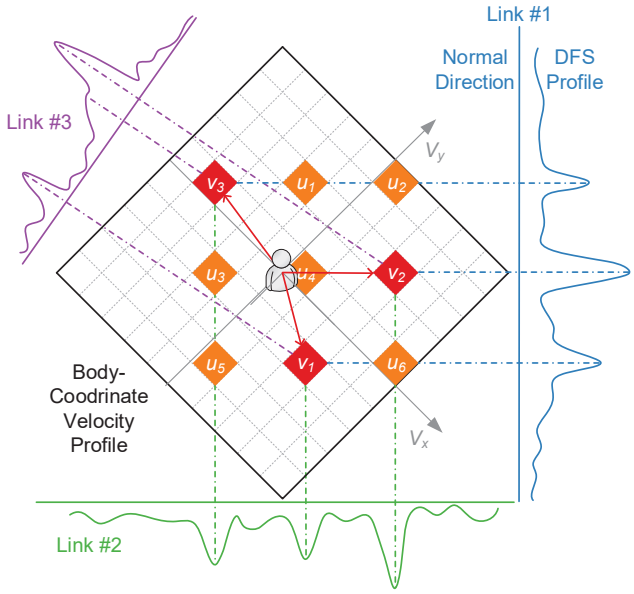


Figure 6: Relationship between the BVP and DFS profiles. Each velocity component in BVP is projected onto the normal direction of a link, and contributes to the power of the corresponding radial velocity component in the DFS profile.

frequency f of subcarriers:

$$\hat{H}(f,t) = \Big(\sum_{l=1}^{L} \alpha_l(f,t) e^{-j2\pi f \tau_l(f,t)}\Big) e^{j\epsilon(f,t)}, \quad (1)$$

where L is the number of paths, α_l and τ_l are the complex attenuation and propagation delay of the l-th path, and $\epsilon(f,t)$ is the phase error caused by timing alignment offset, sampling frequency offset and carrier frequency offset.

By representing phases of multipath signals with the corresponding DFS, CSI can be transformed as [32]:

$$\hat{H}(f,t) = \Big(H_s(f) + \sum_{l \in P_d} \alpha_l(t) e^{j2\pi \int_{-\infty}^{t} f_{D_l}(u)\mathrm{d}u}\Big) e^{j\epsilon(f,t)}, \quad (2)$$

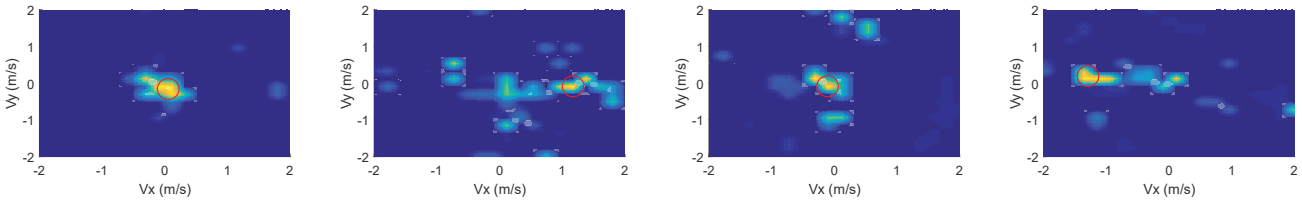
where the constant H_s is the sum of all static signals with zero DFS (e.g., LoS signal), and P_d is the set of dynamic signals with non-zero DFS (e.g., signals reflected by the target).

With conjugate multiplication of CSI of two antennas on the same Wi-Fi NIC calculated, and out-band noises and quasi-static offsets filtered out, random offsets can be removed and only prominent multipath components with non-zero DFS are retained [26]. Further applying short-term Fourier transform yields power distribution over the time and Doppler frequency domains. One example of the spectrogram of a single link is shown in Figure 3. We denote each time snapshot in spectrograms as a DFS profile. Specifically, a DFS profile D is a matrix with dimension as $F \times M$, where F is the number of sampling points in the frequency domain, and M is the number of transceiver links. Based on DFS profile from multiple links, we then derive domain-independent BVP.

4.2 From DFS to BVP

When a person performs a gesture, his body parts (e.g., two hands, two arms and the torso) move at different velocities. As a result, signals reflected by these body parts experience various DFS, which are superimposed at the receiver and form the corresponding DFS profile. As discussed in § 2, while DFS profile contains the information of the gesture, it is also highly specific to the domain. In contrast, the power distribution over physical velocity in the body coordinate system of the person, is only related to the characteristics of the gesture. Thus, in order to remove the impact of domain, BVP is derived out of DFS profiles.

The basic idea of BVP is shown in Figure 6. For practicality, a BVP V is quantized as a discrete matrix with dimension as $N \times N$, where N is the number of possible values of velocity components decomposed along each axis of the body coordinates. For convenience, we establish the local body coordinates whose origin is the location of the person and positive x-axis aligns with the orientation of the person. We will discuss approaches of estimating a person's location and orientation in § 4.4. Currently, it is assumed that the global location and orientation of the person are available. Then the known global locations of wireless transceivers can be transformed into the local body coordinates. Thus, for better clarity, all locations and orientations used in the following derivation are in the local body coordinates. Suppose the locations of the transmitter and the receiver of the i-th link are $\vec{l}_t^{(i)} = (x_t^{(i)}, y_t^{(i)})$, $\vec{l}_r^{(i)} = (x_r^{(i)}, y_r^{(i)})$, respectively, then any velocity components $\vec{v} = (v_x, v_y)$ around the human body (i.e. the origin) will contribute its signal power to some frequency component, denoted as $f^{(i)}(\vec{v})$, in the DFS profile



(a) Stage 1: Start (b) Stage 2: Pushing (c) Stage 3: Stop (d) Stage 4: Pulling

Figure 7: The BVP series of a pushing and pulling gesture. The main velocity component corresponding to the person's hand is highlighted with red circles in all snapshots.

of the i-th link [32]:

$$f^{(i)}(\vec{v}) = a_x^{(i)} v_x + a_y^{(i)} v_y. \tag{3}$$

$a_x^{(i)}$ and $a_y^{(i)}$ are coefficients determined by locations of the transmitter and the receiver:

$$\begin{aligned} a_x^{(i)} &= \frac{1}{\lambda}\left(\frac{x_t^{(i)}}{\|\vec{l}_t^{(i)}\|_2} + \frac{x_r^{(i)}}{\|\vec{l}_r^{(i)}\|_2}\right), \\ a_y^{(i)} &= \frac{1}{\lambda}\left(\frac{y_t^{(i)}}{\|\vec{l}_t^{(i)}\|_2} + \frac{y_r^{(i)}}{\|\vec{l}_r^{(i)}\|_2}\right), \end{aligned} \tag{4}$$

where λ is the wavelength of Wi-Fi signal. As static components with zero DFS (e.g., the line of sight signals and dominant reflections from static objects) are filtered out before DFS profiles are calculated, only signals reflected by the person are retained. Besides, when the person is close to the Wi-Fi link, only signals with one time reflection have prominent magnitudes [33] as Figure 3 shows. Thus, Equation 3 holds valid for the gesture recognition scenario. From the geometric view, Equation 3 means that the 2-D velocity vector $\vec{v}$ is projected on a line whose direction vector is $d^{(i)} = (-a_y^{(i)}, a_x^{(i)})$. Suppose the person is on an ellipse curve whose foci are the transmitter and the receiver of the i-th link, then $d^{(i)}$ is indeed the normal direction of the ellipse at the person's location. Figure 6 shows an example where the person generates three velocity components $\vec{v}_j, j = 1, 2, 3$, and projection of the velocity components on the DFS profiles of three links.

Since coefficients $a_x^{(i)}$ and $a_y^{(i)}$ only depend on the location of the i-th link, the relation of projection of the BVP on the i-th link is fixed. Specifically, an assignment matrix $A^{(i)}_{F\times N^2}$ can be defined:

$$A^{(i)}_{j,k} = \begin{cases} 1 & f_j = f^{(i)}(\vec{v}_k) \\ 0 & \text{else} \end{cases}, \tag{5}$$

where f_j is the j-th frequency sampling point in the DFS profile, and $\vec{v}_k$ is velocity component corresponding to the k-th element of the vectorized BVP V. Thus, the relation between DFS profile of the i-th link and the BVP can be modeled as:

$$D^{(i)} = c^{(i)} A^{(i)} V \tag{6}$$

where $c^{(i)}$ is the scaling factor due to propagation loss of the reflected signal.

4.3 BVP Estimation

How to recover BVP from DFS profiles of only several wireless links is another main challenge because the kinetic profile of a single gesture has hundreds of variables, posing the BVP estimation from DFS profiles as a severely under-determined problem with only a limited number of constraints provided by several wireless links. Specifically, in practice, we estimate one BVP from DFS profiles calculated from 100 ms CSI data. Due to the uncertainty principle, the frequency resolution of DFS profiles is only about 10 Hz. Given that the range of human-induced DFS is within ± 60 Hz [44], the DFS profile of one link can only provide about 12 constraints. In contrast, we moderately set the range and the resolution of velocities along two axes of the body coordinates as ± 2 m/s and 0.2 m/s, respectively, leading to as much as 400 variables! Fortunately, when a person performs a gesture, only a few dominant distinct velocity components exist, due to the limited number of major reflecting multipath signals. Thus, there is an opportunity to correctly recover the BVP from DFS profiles of only several links.

Before a proper solution of BVP developed, it is necessary to understand the minimum number of links required to uniquely recover the BVP. Figure 6 shows an intuitive example with three velocity components $v_j, j = 1, 2, 3$. With only the first two links (blue and green), the three velocity components create three power peaks in each DFS profile. However, when we recover the BVP, there are 9 candidates of velocity components, i.e. $v_j, j = 1, 2, 3$ and $u_k, k = 1, \cdots, 6$. And one can easily find an alternate solution, i.e. $\{u_1, u_3, u_6\}$, meaning that two links are insufficient.

By adding the third link (purple), it is able to resolve the ambiguity with high probability no matter how many velocity components exist, if no overlap of projections happens in the third DFS profile. When projections overlap, however, it is possible that adding the third or even more links cannot resolve the ambiguity. For example, suppose the third link in the Figure 6 is in parallel with the y-axis, and there are three overlaps of projections (i.e. $\{u_1, v_2\}$, $\{v_3, u_4, u_6\}$ and $\{u_3, v_1\}$), then the ambiguous solution $\{u_1, u_3, u_6\}$ is still not resolvable. However, such ambiguity can hardly happen due to its stringent requirement on the distribution of velocity components as well as the orientation of the links. Moreover, we can further reduce the probability of the ambiguity by adding more links. We evaluate the impact of the number of links used by Widar3.0 on system performance in Section 6.5.

With observing of the sparsity of BVP and validating the feasibility of recovering BVP from multiple links, we adopt the idea of

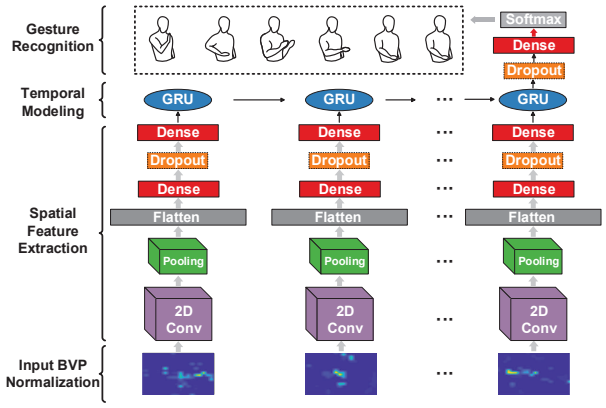


Figure 8: Structure of gesture recognition model.

compressed sensing [13] and formulate the estimation of BVP as an l_0 optimization problem:

$$\min_V \sum_{i=1}^{M} |\text{EMD}(A^{(i)}V, D^i)| + \eta \|V\|_0, \tag{7}$$

where M is the number of Wi-Fi links. The sparsity of the number of the velocity components is coerced by the term $\eta\|V\|_0$, where η represents the sparsity coefficients and $\|\cdot\|_0$ is the number of non-zero velocity components.

EMD$(\cdot,\cdot)$ is the Earth Mover's Distance [35] between two distributions. The selection of EMD rather than Euclidean distance is mainly due to two reasons. First, the quantization of BVP introduces approximation error, i.e. projection of velocity components to the DFS bin might be adjacent to the true one. Such quantization error can be relieved by EMD, which takes the distance between bins into consideration. Second, there are unknown scaling factors between the BVP and DFS profiles, making the Euclidean distance inapplicable.

Figure 7 shows an example of solved BVP series of a pushing and pulling gesture. The dominant velocity component from the hand and the coupling ones from the arm can be clearly observed.

4.4 Location and Orientation Prerequisites

Widar3.0 requires the location and orientation of the person to calculate the domain-independent BVP. In common application scenarios of Widar3.0, when a person wants to interact with the device, he or she approaches it and performs interactive gestures for recognition and response. The antecedent movement of the person gives the chance for estimating his location and orientation, which are the location and moving direction of the person at the end of the trace. Since Wi-Fi based passive tracking has been extensively studied, Widar3.0 can exploit existing sophisticated passive tracking systems, e.g., LiFS [41], IndoTrack [26] and Widar2.0 [33], to obtain the location and orientation of the person. However, Widar3.0 differs from these passive tracking approaches by estimating BVP rather than main torso velocity, and thus further extends the scope of Wi-Fi based sensing. Note that the state-of-the-art localization errors are within several decimeters, and orientation estimation errors are within 20 degrees. We evaluate the impact of location and orientation error by experiments in Section 6.5.

5 RECOGNITION MECHANISM

In Widar3.0, we design a DNN learning model to mining the spatial-temporal characteristics of the BVP series. Figure 8 illustrates the overall structure of the proposed learning model. Specifically, the BVP series is first normalized to remove irrelevant variations caused by instances, persons and hardware settings (§ 5.1). The normalized output is then input into a hybrid deep learning model, which from bottom to top consists of a convolutional neural network (CNN) for spatial feature extraction (§ 5.2) and a recurrent neural network (RNN) for temporal modeling (§ 5.3).

The designed model is a result of the effectiveness of the domain-independent feature BVP. With BVP as input, the hybrid CNN-RNN model can achieve accurate cross-domain gesture recognition although the learning model itself does not possess generalization capabilities. We will verify that the CNN-RNN model is a simple but effective method in Section 6.4.

5.1 BVP Normalization

While BVP is theoretically only related to gestures, two practical factors may affect its stability as the gesture indicator. First, the overall power of BVP may vary due to the adjustment of transmission power. Second, in practice, instances of the same type of gesture performed by different persons may have different time length and moving velocities. Moreover, even instances performed by the same person may slightly vary. Thus, it is necessary to remove these irrelevant factors to retain the simplicity of the learning model.

For signal power variation, Widar3.0 normalizes the element values in each single BVP by adjusting the sum of all elements in BVP to 1. For instance variation, Widar3.0 normalizes the BVP series along the time domain. Specifically, Widar3.0 first sets the standard time length of gestures, denoted as t_0. Then, for a gesture with time length as t, Widar3.0 scales its BVP series to t_0. The assumption behind the scaling operation is that the total distance moved by each body part remains fixed. Thus, to change the time length of the BVP series, Widar3.0 first scales coordinates of all velocity component in the BVP by a factor of $\frac{t}{t_0}$, and then resamples the series to the sampling rate of the original BVP series. After normalization, the output becomes related to gestures only, and is input to the deep learning model.

5.2 Spatial Feature Extraction

The input of the learning model, BVP data, is similar to a sequence of images. Each single BVP describes the power distribution over physical velocity during a sufficiently short time interval. And the continuous BVP series illustrates how the distribution varies corresponding to a certain kind of action. Therefore, to fully understand the derived BVP data, it is intuitive to extract spatial features from each single BVP first and then model the temporal dependencies of the whole series.

CNN is a useful technique to extract spatial features and compress data [27, 47], and it is especially suitable for handling the single BVP, which is highly sparse but preserves spatial locality, as

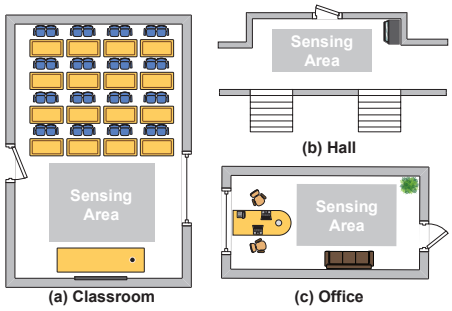


Figure 9: Layouts of three evaluation environments.

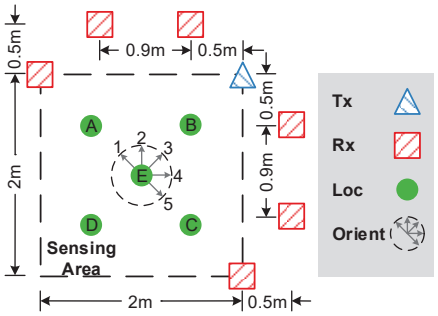


Figure 10: A typical setup of devices and domains in one environment.

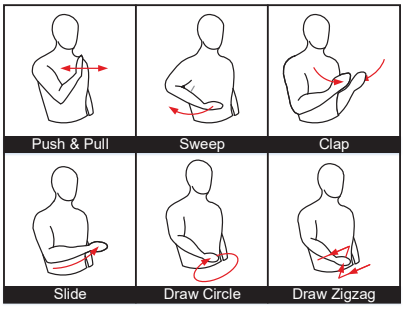


Figure 11: Sketches of gestures evaluated in the experiment.

a velocity component usually corresponds to the same body part as its neighbors with similar velocities. Specifically, the input BVP series, denoted as V, is a tensor with dimension as $N \times N \times T$, where T is the number of BVP snapshots. For the t-th sampling BVP, the matrix $V_{..t}$ is fed into the CNN. Within the CNN, 16 2-D filters are first applied to $V_{..t}$ to obtain local patterns in the velocity domain, which form the output $V^{(1)}_{..t}$. Then, max pooling is applied to $V^{(1)}_{..t}$ to down-sample the features and the output is denoted as $V^{(2)}_{..t}$. With $V^{(2)}_{..t}$ flattened into the vector $\vec{v}^{(2)}_{..t}$, two 64-unit dense layers with ReLU as activation functions are used to further extract features in a higher level. Note that one extra dropout layer is added between two dense layers to reduce overfitting. The final output $\vec{v}_{..t}$ characterizes the t-th sampling BVP. And the output series is used as the input of following recurrent layers for temporal modeling.

5.3 Temporal Modeling

Besides local spatial features within each BVP, BVP series also contains temporal dynamics of the gesture. Recurrent neural networks (RNN) are appealing in that they can model complex temporal dynamics of sequences. There are different types of RNN units, e.g., SimpleRNN, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [12]. Compared with original RNNs, LSTMs and GRUs are more capable of learning long-term dependencies, and we choose GRUs because GRU achieves performance comparable to that of LSTM on sequence modeling, but involves fewer parameters and is easier to train with less data [12].

Specifically, Widar3.0 chooses single-layer GRUs to model the temporal relationships. Inputs $\{\vec{v}_{..t}, t = 1, \cdots, T\}$ output from CNN are fed into GRUs and a 128-dimensional vector $\vec{v}_{..r}$ is generated. Furthermore, a dropout layer is added for regularization, and a softmax classifier with cross-entropy loss for category prediction is utilized. Note that for recognition systems which involve more sophisticated activities with longer durations, the GRU-based models can be transformed into more complex versions [11, 47]. In § 6.4, we will verify that single-layer GRUs are sufficient for capturing temporal dependencies for short-time human gestures.

6 EVALUATION

This section presents the implementation and detailed performance of Widar3.0.

6.1 Experiment Methodology

Implementation. Widar3.0 consists of one transmitter and at least three receivers. All transceivers are off-the-shelf mini-desktops (physical size 170mm × 170mm) equipped with an Intel 5300 wireless NIC. Linux CSI Tool [18] is installed on devices to log CSI measurements. Devices are set to work in the monitor mode, on channel 165 at 5.825 GHz where there are few interfering radios as interference does pose severe impacts on the collected CSI measurements [54]. The transmitter activates one antenna and broadcasts Wi-Fi packets at a rate of 1,000 packets per second. The receiver activates all three antennas which are placed in a line. We implement Widar3.0 in MATLAB and Keras [10].

Evaluation setup. To fully explore the performance of Widar3.0, we conduct extensive experiments on gesture recognition in 3 indoor environments: an empty classroom furnished with desks and chairs, a spacious hall and an office room with furniture like sofa and tables. Figure 9 illustrates the general environmental features and the sensing area in different rooms. Figure 10 shows a typical example of the deployment of devices and domain configurations in the sensing area, which is a 2m × 2m square. Note that the 2m × 2m square is a typical setting to perform interactive gestures for recognition and response, especially in the scenario of smart home, with more Wi-Fi nodes incorporated into smart devices (e.g., smart TV, Xbox Kinect, home gateways, smart camera) to help. We assume that only the gesture performer is in the sensing area as moving entities introduce noisy reflection signals and further result in less accurate DFS profiles of the target gestures. Except for the two receivers and one transmitter placed at the corner of the sensing area, the remaining four receivers can be deployed at random locations outside two sides of the sensing area. As Section 4.3 has mentioned, the deployment of devices hardly pose impacts on Widar3.0 theoretically. All devices are held up at the height of 110 cm, where users with different heights can perform gestures comfortably. In total, 16 volunteers (12 males and 4 females) with different heights (varying from 185 cm to 155 cm) and somatotypes participate in experiments. The ages of the volunteers vary from 22 to 28. And the details of the volunteer information are illustrated in Figure 12.

Dataset. We collect gesture data from 5 locations and 5 orientations in each sensing area, as illustrated in Figure 10. All experiments are approved by our IRB. Two types of datasets are

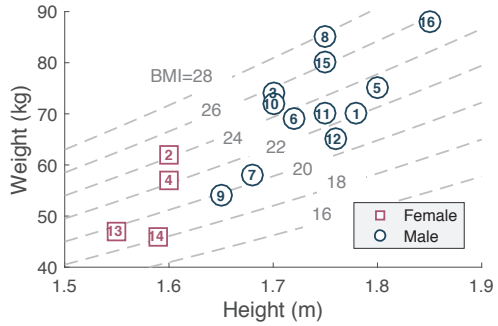


Figure 12: Statistics of participants.

collected. Specifically, the first dataset consists of common hand gestures used in human-computer interaction, including pushing and pulling, sweeping, clapping, sliding, drawing circle and drawing zigzag. The sketches of the six gestures are plotted in Figure 11. This dataset contains 12,000 gesture samples (16 users × 5 positions × 5 orientations × 6 gestures × 5 instances). The second dataset is collected for a case study of more complex and semantic gestures. Two volunteers (one male and one female) draw number 0 ∼ 9 in the horizontal plane, and totally 5,000 samples (2 users × 5 positions × 5 orientations × 10 gestures × 10 instances) are collected. Before collecting the datasets, we ask volunteers to watch the example video of each gesture. The datasets and the example videos are available at website[1].

Prerequisites Acquisition. The position and orientation of the user are prerequisites for calculation of BVP. In general, the last estimation of location and the last estimation of moving direction can be provided by tracking systems[26, 33, 41], as the location and orientation of the user in Widar3.0. Note that the function of Widar3.0 is independent of that of the motion tracking system. To fully understand how Widar3.0 works, we record the ground truth of location and orientation of the user in most experiments, and explicitly introduce location and orientation error in the parameter study (Section 6.5) to evaluate the relation between recognition accuracy and location and orientation errors.

6.2 Overall Accuracy

Taking all domain factors into consideration, Widar3.0 achieves an overall accuracy of 92.7%, with 90 and 10 percentage data collected in Room 1 used for training and testing, respectively. Figure 13a shows the confusion matrix of 6 gestures in dataset 1, and Widar3.0 achieves consistently high accuracy of over 85% for all gestures. We also conduct experiments with gestures of an "unknown" class are additionally added. Volunteers are required to perform arbitrary gestures except for the above 6 gestures. The overall accuracy drops to 90.1% and Widar3.0 can differentiate the unknown class with an accuracy of 87.1%. The reasons are as follows. On one hand, gestures from an "unknown" class might be similar to the predefined ones to a certain degree. On the other hand, the collected "unknown" gestures are still limited. We believe the results can be further improved if we introduce additional filtering mechanisms or modify the learning model to solve the issue of "novelty detection", which is another significant topic in recognition problems.

Figure 13b, 13c, 13d and 13e further show confusion matrices considering each specific domain factors. For each domain factor, we calculate average accuracy of cases where one out of all domain instances are used for testing, while the rest domain instances are for training. The average accuracy over all gestures are provided as well, and it can be seen that Widar3.0 achieves consistent high performance across different domains, demonstrating its capability of cross-domain recognition.

We observe that for both in-domain and cross-domain cases, the gestures "pushing and pulling", "drawing circle" and "drawing zigzag" usually correspond to a lower accuracy. While the "pushing and pulling" gesture is the simplest one among all gestures, it is performed just in front of the user torso, and is more likely to be blocked from the perspectives of certain links, which results in less accurate BVP estimation as shown in the following experiments (Section 6.5). When users perform the gesture "drawing circle" or "drawing zigzag", the trajectory has significant changes in vertical direction. However, Widar3.0 is designed to extract BVP only in the horizontal plane, leading to information loss for the two gestures, and decrease in recognition accuracy.

Case study. We now examine if Widar3.0 still works well for more complex gesture recognition tasks. In this case study, volunteers draw number 0∼9 in the horizontal plane and 5,000 samples are collected in total. We divide the dataset into training and testing randomly with the ratio 9:1. As shown in Figure 13f, Widar3.0 achieves satisfying results of over 90% for 8 gestures and the average accuracy is 92.9%.

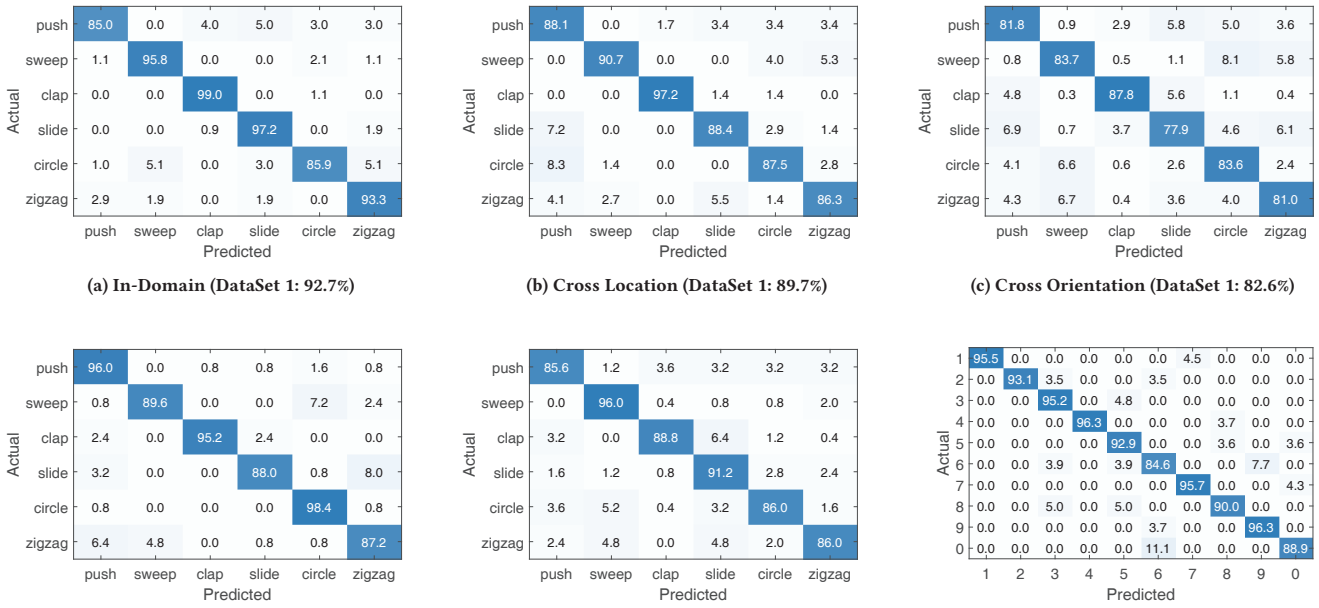
6.3 Cross-Domain Evaluation

We now evaluate the overall performance of Widar3.0 on across different domain factors, including environment, person diversity and location and orientation of the person. For evaluation on each domain factor, we keep the other domain factors unchanged, and perform leave-one-out cross validation on the datasets. The system performance, in terms of mean and variance of the accuracy, is shown in Figure 14∼17.

Location independence. The model is trained on the BVPs of random 4 locations, all 5 orientations and 8 people in Room 1. And the data collected at the last location in the same room is used for testing. As shown in Figure 14, the average accuracies for all locations uninvolved in training are all above 85%. Widar3.0 achieves best performance of 92.3% with location e, which is at the center of the sensing area, as the target domain. The accuracy descends to 85.3% when testing dataset is collected at location d, as wireless signal reflected by human-body becomes weaker after a longer distance of propagation, which leads to less accurate BVPs. In addition, BVP is modeled from signals reflected by the person. If the person happens to pass his arm through the line-of-sight path of any links, the accuracy will slightly drop, as proved by the result of location b.

Orientation sensitivity. In this experiment, we select each orientation as the target domain and other 4 orientations as the source domain. Figure 15 shows that the accuracy remains above 80% for orientation 2, 3, 4. Compared with best target orientation 3, whose

[1] http://tns.thss.tsinghua.edu.cn/widar3.0/index.html



(a) In-Domain (DataSet 1: 92.7%) **(b) Cross Location (DataSet 1: 89.7%)** **(c) Cross Orientation (DataSet 1: 82.6%)**

(d) Cross Environment (DataSet 1: 92.4%) **(e) Cross Person (DataSet 1: 88.9%)** **(f) In-Domain (DataSet 2: 92.9%)**

Figure 13: Confusion matrices of different settings with two gesture datasets.

accuracy is around 90%, the performance at orientation 1&5 declines by over 10%. The reason is that gestures might be shadowed by human body in these two orientations and the number of effective wireless links for BVP generation decreases. For common gesture recognition applications (e.g., TV control), however, it is reasonable to assume that when the user faces towards the TV, his orientation does not deviate much from most wireless devices, a sufficient number of which could be used for accurate gesture recognition.

Environment diversity. The accuracy across different environments is another significant criterion for performance of cross-domain recognition. In this experiment, gesture samples collected in room 1 are used as the training dataset, and three groups of gesture samples collected in three rooms are used as testing datasets. As Figure 16 depicts, while the accuracy for different rooms slightly drops, the average accuracy preserves over 87% even if the environment changes totally. In a nutshell, Widar3.0 is robust to different environments.

Person variety. Data collected from different persons may have discrepancy due to their various behavior patterns. Widar3.0 incorporates BVP normalization to alleviate this problem. To evaluate the performance of Widar3.0 on different users, we train the model on a dataset from every combination of 7 persons, and then test with the data of the resting person. Figure 17 shows that the accuracy remains over 85% across 7 persons. The impact of the number of persons used in training the recognition model is further investigated in Section 6.5.

6.4 Method Comparison

This section compares the capability of cross-domain recognition with different methods, learning features and structures of learning networks. In the experiment, training and testing datasets are collected separately in Room 1 and 2.

Comparison with the state-of-the-arts works. We compare Widar3.0 against several alternative state-of-the-arts methodologies, CARM[44], EI[20] and CrossSense[50], where the latter two are feasible for cross-domain recognition. Specifically, CARM uses DFS profiles as learning features and adopts HMM model. EI incorporates an adversarial network and specializes the training loss to additionally exploit characteristics of unlabeled data in target domains. CrossSense proposes an ANN-based roaming model to translate signal features from source domains to target domains, and employs multiple expert models for gesture recognition. Figure 18 shows the system performance of the four approaches. Widar3.0 achieves better performance with the state-of-the-art cross-domain learning methodologies, EI and CrossSense, and it does not require extra data from a new domain or model re-training. In contrast, both feature and learning model of CARM do not have cross-domain capability, which is the main reason for its significantly lower recognition accuracy.

Comparison of input features. We compare three types of features with different levels of abstraction from raw CSI measurements, i.e. denoised CSI, DFS profiles and BVP, by feeding them into the CNN-GRU hybrid deep learning model, similar to that in Widar3.0. Specifically, the size of denoised CSI is 18 (the number of antennas of 6 receivers) $\times$ 30 (the number of subcarriers) $\times$ T (the number of time samples), and the DFS profile has the shape as 6

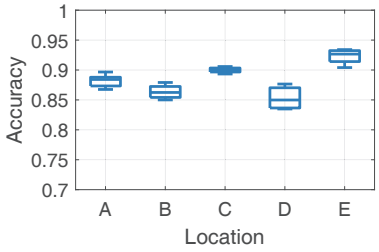


Figure 14: Accuracy distributions for cross-location evaluation.

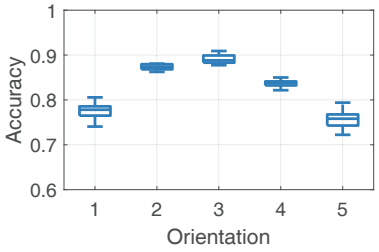


Figure 15: Accuracy distributions for cross-orientation evaluation.

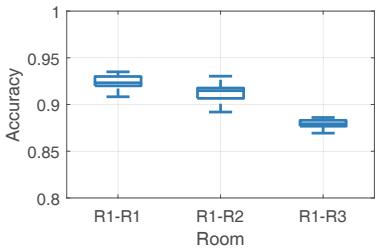


Figure 16: Accuracy distributions for cross-environment evaluation.

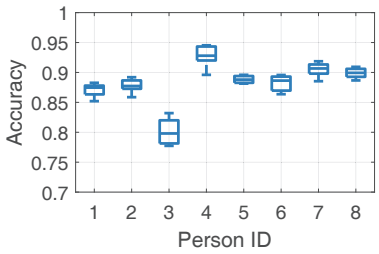


Figure 17: Accuracy distributions for cross-person evaluation.

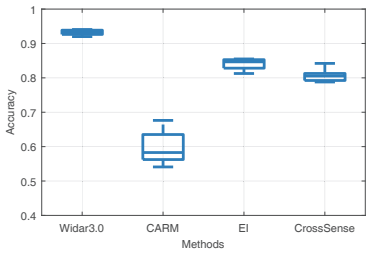


Figure 18: Comparison of recognition approaches.

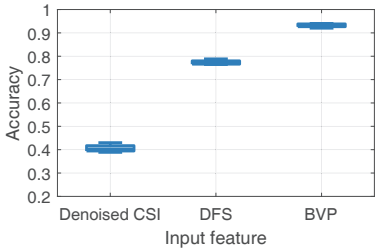


Figure 19: Comparison of input features.

(the number of receivers) × F (the number of Doppler frequency samples) × T (the number of time samples). As shown in Figure 19, BVP outperforms both denoised CSI and DFS, with an increase of accuracy by 52% and 15%, respectively. The performance improvement of BVP attributes its immunity to changes of layouts of transceivers, which however may significantly influences the other two types of features.

Comparison of learning model structures. Different deep learning models are further compared and the system performance is demonstrated in Figure 20. Specifically, the CNN-GRU hybrid model increases the accuracy by around 5% compared with the simple GRU model which merely captures temporal dependencies. The former model benefits from representative high-level spatial features within each BVP snapshot. In addition, we also feed BVP into a two-convolutional-layer CNN-GRU hybrid model and a CNN-Hierarchical-GRU model [11]. It is shown that a more complex deep learning model does not promote the performance, demonstrating that BVP of different gestures are distinct enough to be discriminated by a simple but effective classifier.

6.5 Parameter Study

Impact of link numbers. In the above experiments, 6 links are deployed for more accurate estimation of BVP. This section studies the impact of the number of links on system performance. As shown in Figure 21, the accuracy gradually decreases as the number of links reduces from 6 to 3, but experiences a more significant drop when only two links are used. The main reason is that some BVPs cannot be correctly recovered with only 2 links considering the ambiguity mentioned in Section 4.3, and gestures at certain locations or orientations cannot be fully captured due to blockage.

Impact of location and orientation estimation error. Localizations and orientations provided by Wi-Fi based motion tracking systems usually have errors of about several decimeters and 20 degrees, respectively. Thus, it is necessary to understand how these errors impact the performance of Widar3.0. Specifically, we record ground truth of location and orientation, and calculate errors where gestures are performed. On one hand, as shown in Figure 22, the overall accuracy remains over 90% when the location error is within 40 cm, but then drops as the error further increases. On the other hand, Figure 23 shows that the overall accuracy gradually drops with more deviation of orientation. While the tracking errors negatively impact the performance of Widar3.0, taking practical location and orientation errors into consideration, we believe existing motion tracking works can still provide location and orientation results with acceptable accuracy.

Impact of training set diversity. This experiment studies how the number of volunteers in training dataset impacts the performance. Specifically, a varying number of volunteers from 1 to 7 participate in collecting the training dataset, and data from another new person is used to test Widar3.0. Figure 24 shows that the average gesture recognition accuracy decreases from 89% to 74% when the number of people for training varies from 7 to 1. The reasons come from two folds. First, with the training dataset contributed by fewer volunteers, the deep learning model will be less thoroughly trained. Second, the behavior difference between testing persons and training persons will be amplified even if we have adopted BVP normalization. In general, Widar3.0 promises an accuracy of over 85% with more than 4 people in the training set.

Impact of transmission rates. As Widar3.0 requires packet transmission for gesture recognition, normal communication flow

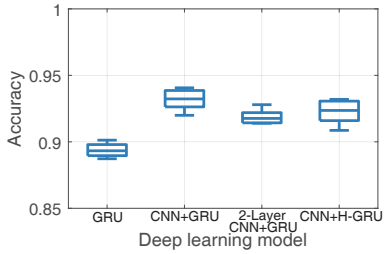


Figure 20: Comparison of DNNs.

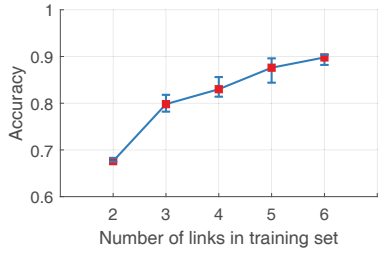


Figure 21: Impact of link numbers.

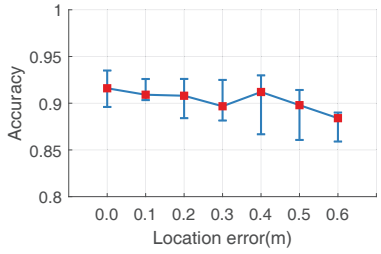


Figure 22: Impact of location error.

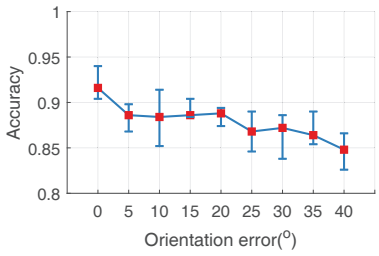


Figure 23: Impact of orientation error.

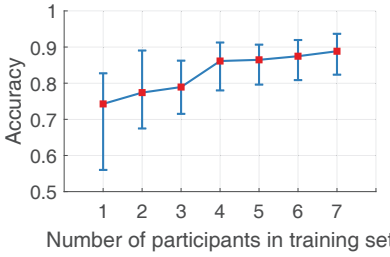


Figure 24: Impact of training diversity.

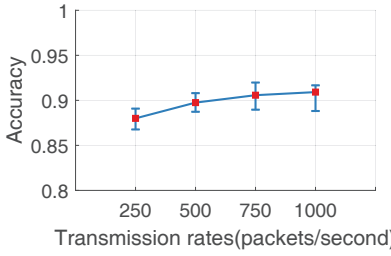


Figure 25: Impact of transmission rates.

might be interfered. Therefore, we evaluate the performance of Widar3.0 with different CSI transmission rates. We collect CSI measurements at the initial transmission rate of 1,000 packets per second, and down-sample the CSI series to 750 Hz, 500 Hz, 250 Hz. Figure 25 shows that the accuracy degrades slightly by around 4% when the sampling rate drops to 250Hz, and remains over 85% for all cases. In addition, Widar3.0 can further reduce the impacts on communication with shorter packets used as only CSI measurements are useful for the recognition tasks.

7 DISCUSSIONS

User height. Since transceivers are placed at the same height, CSI measurements mainly capture the horizontal velocity components. Thus, different user heights may impact the recognition performance of Widar3.0, as the devices may observe different groups of velocity components intercepted at this height. However, Widar3.0 still has the capability of recognizing gestures in 3-D space, as common gestures remain their uniqueness even within the fixed height. As shown in the experiments, Widar3.0 is able to recognize gestures "draw circle" and "draw zigzag", which both contain vertical velocity components due to the fixed length of arms. By regarding the person as on an ellipsoid whose foci are the transceivers of a link, the BVP can be further generalized to 3-D space. Further work includes optimizing the deployment of Wi-Fi links to enable calculation of 3-D BVP, and revising the learning model with 3-D BVPs as input.

Number of Wi-Fi links for gesture recognition. Although three wireless links are sufficient to resolve the ambiguity with high probability for BVP generation, six receivers in total are deployed in the experiments. The reasons are two folds. First, compared with macro activities, the reflected signal of micro gestures is much weaker, since the effective area of hand and arm is much smaller than that of torso and leg, resulting in less prominent DFS profiles. Second, gestures with hands and arms may be opportunistically shadowed by other body parts when the user faces away from the link. For macro activity such as walking, running, jumping and falling, it is believed that the number of Wi-Fi links required for recognition can be reduced. It is worth noting that Widar3.0 does not require fixed deployment of Wi-Fi devices in the environment, as BVP is the power distribution over absolute velocities.

Applications beyond gesture recognition. While Widar3.0 is a Wi-Fi based gesture system, the feature used in Widar3.0, BVP, can theoretically capture movements over the whole body of the person, and thus is envisioned to be used in other device-free sensing scenarios, such as macro activity recognition, gait analysis and user identification. In these scenarios where users are likely to continuously change their locations and orientations, BVP calculation and motion tracking approaches can be intermittently invoked to obtain BVPs along the whole trace, which then may serve as a unique indicator for user's activity or identity.

8 RELATED WORK

Our work is highly related to wireless human sensing techniques, which are roughly categorized into model-based and learning-based ones, targeting at localization and activity recognition, respectively.

Model-based wireless localization. Model-based human sensing explicitly builds physical link between wireless signals and human movements. On the signal side, existing approaches extract various parameters of signals reflected or shadowed by human, including DFS [26, 32, 44], ToF [2–4, 21], AoA/AoD [2, 5, 21, 25] and attenuation [7, 41]. Based on types of devices used, parameters with different extent of accuracy and resolution can be obtained. WiTrack [3, 4] develops FMCW radar with wide bandwidth to accurately estimate ToFs of reflected signals. WiDeo [21] customizes

full-duplex Wi-Fi to jointly estimate ToFs and AoAs of major reflectors. In contrast, though limited by the bandwidth and antenna number, Widar2.0 [33] improves resolution by jointly estimating ToF, AoA and DFS.

On the human side, existing model-based works only tracks coarse human motion status, such as location [4, 41], velocity [26, 32], gait [43, 49] and figure [2, 19]. Though not detailed enough, they provide coarse human movement information, which can further help Widar3.0 and other learning-based activity recognition works to remove domain dependencies of input signal features.

Learning-based wireless activity recognition. Due to complexity of human activity, existing approaches extract signal features, either statistical [14, 15, 23, 28, 30, 45, 49] or physical [6, 31, 34, 38, 39, 44, 51, 52] ones, and map them to discrete activities. The statistical methods treat the wireless signal as time series data, extract its waveforms and distributions in both time and frequency domain as fingerprints. E-eyes [45] is a pioneer work to use strength distribution of commercial Wi-Fi signals and KNN to recognize human activities. Niu et al. [30] uses signal waveforms for fine-grained gesture recognition. The physical methods take a step further to extract features with clear physical meanings. CARM [44] calculates power distribution of DFS components as learning features of HMM model. WIMU [38] further segments DFS power profile for multi-person activity recognition. However, due to fundamental limits of domain dependencies of wireless signals, directly using either statistical or physical features is infeasible to generalize to different domains.

Tempts to adapt recognition schemes in various domains fall into two categories: virtually generating features for target domains [39, 40, 50, 53] and developing domain-independent features [9, 20, 37]. In the former type, WiAG [39] derives translation functions between CSIs from different domains, and generates virtual training data accordingly. CrossSense [50] adopts the idea of transfer learning, and proposes a roaming model to translate signal features between domains. However, features generated by these types of works are still domain-dependent, which require training of classifier for each individual domain, leading to a waste of training efforts. In contrast, with the help of passive localization, Widar3.0 directly uses domain-independent BVPs as features and trains the classifier only once.

In the latter type, the idea of adversarial learning is usually adopted to shift the task of separating gesture-related features from domain-related ones. EI [20] incorporates an adversarial network to obtain domain-independent features from CSI. However, cross-domain learning methodologies require extra data samples from the target domain, increasing data collection and training efforts. Moreover, features generated by learning models are semantically uninterpretable. In contrast, Widar3.0 explicitly extracts domain-independent BVPs, and only needs a simply designed learning model without the capability of cross-domain learning.

9 CONCLUSION

In this paper, we propose a Wi-Fi based zero-effort cross-domain gesture recognition system. First, we model the quantitative relation between complex gestures and CSI dynamics, and extract velocity profiles of gestures in body coordinates, which are domain-independent and act as unique indicators of gestures. Then, we develop a one-fits-all deep learning model to fully exploit spatial-temporal characteristics of BVP for gesture recognition. We implement Widar3.0 on COTS Wi-Fi devices and evaluate it in real environments. Experimental results show that Widar3.0 achieves high recognition accuracy across different domain factors, specifically, 89.7%, 82.6%, 92.4% and 88.9% for user's location, orientation, environment and user diversity, respectively. Future work focuses on applying Widar3.0 to fortify various sensing applications.

ACKNOWLEDGMENTS

We sincerely thank our shepherd Professor Yingying Chen and the anonymous reviewers for their valuable feedback. We also thank Junbo Zhang, the undergraduate student at Tsinghua University, for helping to build the platform. This work is supported in part by the National Key Research Plan under grant No. 2016YFC0700100, NSFC under grants 61832010, 61632008, 61672319, 61872081, and National Science Foundation under grant CNS-1837146.

REFERENCES

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A Ubiquitous Wifi-based Gesture Recognition System. In *Proceedings of IEEE INFOCOM*. Kowloon, Hong Kong.
[2] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the Human Figure Through a Wall. *ACM Transactions on Graphics* 34, 6 (November 2015), 219:1–219:13.
[3] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. Multi-Person Localization via RF Body Reflections. In *Proceedings of USENIX NSDI*. Oakland, CA, USA.
[4] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 2014. 3d Tracking via Body Radio Reflections. In *Proceedings of USENIX NSDI*. Seattle, WA, USA.
[5] Fadel Adib and Dina Katabi. 2013. See Through Walls with Wi-Fi!. In *Proceedings of ACM SIGCOMM*. Hong Kong, China.
[6] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. 2017. Recognizing keystrokes using WiFi devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (May 2017), 1175–1190.
[7] Maurizio Bocca, Ossi Kaltiokallio, Neal Patwari, and Suresh Venkatasubramanian. 2014. Multiple Target Tracking with RF Sensor Networks. *IEEE Transactions on Mobile Computing* 13, 8 (August 2014), 1787–1800.
[8] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *ACM Comput. Surv.* 46, 3 (January 2014), 33:1–33:33.
[9] Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. 2018. Distributionally Robust Semi-Supervised Learning for People-Centric Sensing. In *Proceedings of AAAI*. New Orleans, LA, USA.
[10] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.
[11] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. 2016. Hierarchical Multiscale Recurrent Neural Networks. *CoRR* abs/1609.01704 (2016).
[12] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014).
[13] David L Donoho. 2006. Compressed Sensing. *IEEE Transactions on Information Theory* 52, 4 (April 2006), 1289–1306.
[14] Biyi Fang, Nicholas D Lane, Mi Zhang, Aidan Boran, and Fahim Kawsar. 2016. BodyScan: A Wearable Device for Contact-less Radio-based Sensing of Body-related Activities. In *Proceedings of ACM MobiSys*. Singapore, Singapore.
[15] Biyi Fang, Nicholas D Lane, Mi Zhang, and Fahim Kawsar. 2016. HeadScan: A Wearable System for Radio-based Sensing of Head and Mouth-related Activities. In *Proceedings of ACM/IEEE IPSN*. Vienna, Austria.
[16] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and Recognizing Human-Object Interactions. In *Proceedings of IEEE CVPR*. Salt Lake City, UT, USA.
[17] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (June 2017), 11:1–11:28.
[18] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool Release: Gathering 802.11n Traces with Channel State Information. *ACM SIGCOMM Computer Communication Review* 41, 1 (January 2011), 53–53.
[19] Donny Huang, Rajalakshmi Nandakumar, and Shyamnath Gollakota. 2014. Feasibility and Limits of Wi-Fi Imaging. In *Proceedings of ACM MobiSys*. Bretton

Woods, NH, USA.
[20] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of ACM MobiCom*. New Delhi, India.
[21] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. 2015. Wideo: Fine-grained Device-free Motion Tracing Using RF Backscatter. In *Proceedings of USENIX NSDI*. Oakland, CA, USA.
[22] Kaustubh Kalgaonkar and Bhiksha Raj. 2009. One-Handed Gesture Recognition Using Ultrasonic Doppler Sonar. In *Proceedings of IEEE ICASSP*. Taipei, Taiwan.
[23] Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: Talk to Your Smart Devices with Finger-grained Gesture. In *Proceedings of ACM UbiComp*. Heidelberg, Germany.
[24] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical Human Sensing in the Light. In *Proceedings of ACM MobiSys*. Singapore, Singapore.
[25] Xiang Li, Shengjie Li, Daqing Zhang, Jie Xiong, Yasha Wang, and Hong Mei. 2016. Dynamic-MUSIC: Accurate Device-Free Indoor Localization. In *Proceedings of ACM UbiComp*. Heidelberg, Germany.
[26] Xiang Li, Daqing Zhang, Qin Lv, Jie Xiong, Shengjie Li, Yue Zhang, and Hong Mei. 2017. IndoTrack: Device-Free Indoor Human Tracking with Commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (September 2017), 72:1–72:22.
[27] Cihang Liu, Lan Zhang, Zongqian Liu, Kebin Liu, Xiangyang Li, and Yunhao Liu. 2016. Lasagna: Towards Deep Hierarchical Understanding and Searching over Mobile Sensing Data. In *Proceedings of ACM MobiCom*. New York City, NY, USA.
[28] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign Language Recognition Using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (March 2018), 23:1–23:21.
[29] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity Information Leakage Using Music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (September 2017), 87:1–87:24.
[30] Kai Niu, Fusang Zhang, Jie Xiong, Xiang Li, Enze Yi, and Daqing Zhang. 2018. Boosting Fine-grained Activity Sensing by Embracing Wireless Multipath Effects. In *Proceedings of ACM CoNEXT*. Heraklion/Crete, Greece.
[31] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-Home Gesture Recognition Using Wireless Signals. In *Proceedings of ACM MobiCom*. Miami, FL, USA.
[32] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-Level Passive Tracking via Velocity Monitoring with Commodity Wi-Fi. In *Proceedings of ACM MobiHoc*. Chennai, India.
[33] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2.0: Passive Human Tracking with a Single Wi-Fi Link. In *Proceedings of ACM MobiSys*. Munich, Germany.
[34] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring Motion Direction Using Commodity Wi-Fi for Interactive Exergames. In *Proceedings of ACM CHI*. Denver, CO, USA.
[35] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 2 (November 2000), 99–121.
[36] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a Smartwatch and I can Track my User's Arm. In *Proceedings of ACM MobiSys*. Singapore, Singapore.
[37] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation. In *Proceedings of ICLR*. Vancouver, Canada.
[38] Raghav H. Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-User Gesture Recognition Using WiFi. In *Proceedings of ACM MobiSys*. Munich, Germany.
[39] Aditya Virmani and Muhammad Shahzad. 2017. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proceedings of ACM MobiSys*. Niagara Falls, NY, USA.
[40] Jindong Wang, Yiqiang Chen, Lisha Hu, Xiaohui Peng, and Philip S Yu. 2017. Stratified Transfer Learning for Cross-domain Activity Recognition. In *Proceedings of IEEE PerCom*. Big Island, HI, USA.
[41] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. 2016. LiFS: Low Human-effort, Device-free Localization with Fine-grained Subcarrier Information. In *Proceedings of ACM MobiCom*. New York City, NY, USA.
[42] Minsi Wang, Bingbing Ni, and Xiaokang Yang. 2017. Recurrent Modeling of Interaction Context for Collective Activity Recognition. In *Proceedings of IEEE CVPR*. Honolulu, HI, USA.
[43] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait Recognition Using WiFi Signals. In *Proceedings of ACM UbiComp*. Heidelberg, Germany.
[44] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2017. Device-Free Human Activity Recognition Using Commercial WiFi Devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (May 2017), 1118–1131.
[45] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: Device-free Location-oriented Activity Identification Using Fine-grained WiFi Signatures. In *Proceedings of ACM MobiCom*. Maui, HI, USA.
[46] Zheng Yang, Zimu Zhou, and Yunhao Liu. 2013. From RSSI to CSI: Indoor Localization via Channel Response. *ACM Comput. Surv.* 46, 2 (November 2013), 25:1–25:32.
[47] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of ACM WWW*. Perth, Australia.
[48] Koji Yatani and Khai N Truong. 2012. BodyScope: A Wearable Acoustic Sensor for Activity Recognition. In *Proceedings of ACM UbiComp*. Pittsburgh, PA, USA.
[49] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. 2016. WiWho: WiFi-Based Person Identification in Smart Spaces. In *Proceedings of ACM/IEEE IPSN*. Vienna, Austria.
[50] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Tapio Nurmi, and Zheng Wang. 2018. CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *Proceedings of ACM MobiCom*. New Delhi, India.
[51] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *Proceedings of IEEE CVPR*. Salt Lake City, UT, USA.
[52] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-Based 3D Skeletons. In *Proceedings of ACM SIGCOMM*. Budapest, Hungary.
[53] Zhongtang Zhao, Yiqiang Chen, Junfa Liu, Zhiqi Shen, and Mingjie Liu. 2011. Cross-People Mobile-Phone Based Activity Recognition. In *Proceedings of IJCAI*. Barcelona, Spain.
[54] Yue Zheng, Chenshu Wu, Kun Qian, Zheng Yang, and Yunhao Liu. 2017. Detecting Radio Frequency Interference for CSI Measurements on COTS WiFi Devices. In *Proceedings of IEEE ICC*. Paris, France.