Towards Domain-independent Complex and Fine-grained Gesture Recognition with RFID

CAO DIAN, Shanghai Jiao Tong University, China
DONG WANG*, Shanghai Jiao Tong University, China
QIAN ZHANG, Shanghai Jiao Tong University, China
RUN ZHAO, Shanghai Jiao Tong University, China
YINGGANG YU, Shanghai Jiao Tong University, China

Gesture recognition plays a fundamental role in emerging Human-Computer Interaction (HCI) paradigms. Recent advances in wireless sensing show promise for device-free and pervasive gesture recognition. Among them, RFID has gained much attention given its low-cost, light-weight and pervasiveness, but pioneer studies on RFID sensing still suffer two major problems when it comes to gesture recognition. The first is they are only evaluated on simple whole-body activities, rather than complex and fine-grained hand gestures. The second is they can not effectively work without retraining in new domains, i.e. new users or environments. To tackle these problems, in this paper, we propose RFree-GR, a domain-independent RFID system for complex and fine-grained gesture recognition. First of all, we exploit signals from the multi-tag array to profile the sophisticated spatio-temporal changes of hand gestures. Then, we elaborate a Multimodal Convolutional Neural Network (MCNN) to aggregate information across signals and abstract complex spatio-temporal patterns. Furthermore, we introduce an adversarial model to our deep learning architecture to remove domain-specific information while retaining information relevant to gesture recognition. We extensively evaluate RFree-GR on 16 commonly used American Sign Language (ASL) words. The average accuracy for new users and environments (new setup and new position) are 89.03%, 90.21% and 88.38%, respectively, significantly outperforming existing RFID based solutions, which demonstrates the superior effectiveness and generalizability of RFree-GR.

CCS Concepts: • **Human-centered computing → Gestural input**; **Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: RFID; device-free; gesture recognition; multimodal data fuse; adversarial learning

ACM Reference Format:
Cao Dian, Dong Wang, Qian Zhang, Run Zhao, and Yinggang Yu. 2020. Towards Domain-independent Complex and Fine-grained Gesture Recognition with RFID. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 187 (November 2020), 22 pages. https://doi.org/10.1145/3427315

*Dong Wang is the corresponding author

Authors' addresses: Cao Dian, sj_cdd@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Dong Wang, wangdong@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Qian Zhang, qwert3472@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Run Zhao, zhaorun@cs.sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Yinggang Yu, yinggang_yu@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China.

2573-0142/2020/11-ART187 $15.00
https://doi.org/10.1145/3427315

1 INTRODUCTION

Recent years there has been an increasing interest in gesture based Human-Computer Interaction. Compared to traditional methods, e.g. typing and touching, gestures can deliver a more natural and intuitive communication with computers. Furthermore, unlike voice input, gestural interface provides a better user experience in noisy environments and more importantly, is readily accessible to the deaf. Therefore, gesture recognition can be a key enabler for a wide range of innovative applications. For example, in public areas such as hospitals, libraries, and supermarkets, gesture recognition enables people to interact with self-service devices in a contactless way, which does not only improve user experience, but also avoid possible bacteria infection through devices. In addition, in home settings, gesture recognition releases people from physical interfaces, such as knobs and levers when fine-tuning appliances. People can also use gestures to interact with smart voice assistants, which brings much convenience to the deaf.

Traditional solutions for gesture recognition mainly rely on cameras [18, 31] or wearable devices [13, 23, 39]. Although achieving remarkable recognition accuracy, these approaches suffer from various limitations such as sensitivity to lighting conditions, leakage of privacy [21, 36] or requirement of on-body sensors. In contrast, the recent wireless sensing techniques provide a more promising alternative for device-free gesture recognition which means the user does not need to wear any devices. The basic idea of wireless sensing is to model and extract motion induced variations on wireless signals to infer gestures [22]. During the past decade, extensive efforts have been made on wireless sensing through various wireless techniques such as WiFi [17, 28, 34], RFID [1, 2, 29, 43, 47], ultrasound [9, 33], and millimeter radar [16, 32]. Among them, RFID has attracted much attention given its low-cost, light-weight, and pervasiveness. However, state-of-the-art RFID sensing researches mainly focus on coarse-grained whole-body activities or gestures like falling, walking, pull or push, while many gesture based applications need to sense complex and fine-grained gestures like finger movements and sign gestures. Moreover, existing solutions only work well when the testing domain appears in the training set, and their performance will degrade dramatically when it comes to a new user or a new environment.

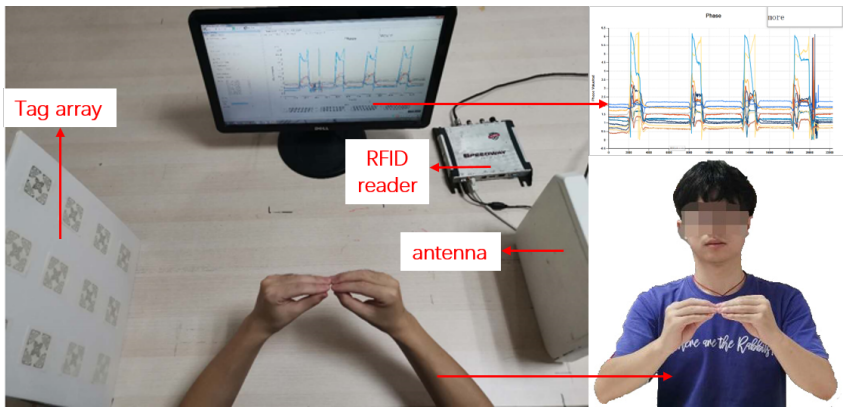


Fig. 1. Scenario of RFree-GR.

To tackle these issues mentioned above, in this paper, we explore the feasibility and mechanisms of RFID for gesture recognition, and present RFree-GR, a device-free system that enables domain-independent complex and fine-grained gesture recognition. As shown in Figure 1, we place a tag array and an antenna face to face. When the user performs gestures between the tag array and

the antenna, RFree-GR can capture the signal variations and map them to the performed gestures. Given that sign language is regarded as the most grammatically structured category of gestural communications, we exploit sign language to design and build our system.

There exist three challenges in designing RFree-GR. The first is **complex and fine-grained gesture sensing**. Sign gestures involve complex, diverse and fine-grained movements. Besides, many gestures share very similar movements. Unfortunately, the signal indicators, i.e. RSS (Received Signal Strength) and phase, provided by commercial RFID readers have limited spatial resolution. So they are far from profiling the complex spatio-temporal changes of gestures, and even worse, due to multipath fading, the sensing ability of RF signals can be very weak for subtle finger movements in some certain positions. The second is **multimodal and multi-tag signal fusion**. The RSS and phase signal of each tag carry information with various perspectives. They often complement each other, allowing for useful information gain. However, as signal of each type has different noise distributions, spatio-temporal representations and correlation structures, building models for signal fusion to fully leverage information contained within and across each type is not trivial but very important for accurate gesture recognition. The third is **domain-independent feature extraction**. The RF signals carry much information that is irrelevant to gestures, and are highly dependent on the gesture performers and around environments which are called domains in our paper. As a result, the gesture recognition model trained in a specific domain usually undergoes drastically drop in performance with a new domain. Domain-independent feature extraction is a very challenging issue in activity recognition field, and the characteristics of RF signals make it more difficult to realize.

In order to solve the above challenges, we are looking for solutions in the context of raw signal and deep learning networks. First, as RFID readers can only provide limited signal information, we elaborates an 2D multi-tag array to enhance the sensing ability. The complementarity among these spatially distributed tags enables RFree-GR to well capture the complex and fine-grained gestures. Besides, we further enhance the sensing capacity of RF signals by removing static components and augmenting data diversity. Second, we design a Multimodal Convolutional Neural Network (MCNN) as the feature extractor to extract spatio-temporal patterns from multimodal signals of each tag. Specifically, we exploit spatio-temporal convolutions to abstract intra-modal features across space and time, and then get the higher-level cross-modal features via a modality merged networks. Third, to improve the generalization performance of RFree-GR to new domains, we introduce an adversarial training regime to our deep learning networks. The adversarial training ensures the feature extractor to discard information specific to domains while retaining information relevant to gesture recognition.

To evaluate RFree-GR, we collect 6080 samples of 16 commonly used American Sign Language (ASL) words from 15 volunteers in five different setups and five different positions. The setup and position refer to the distance between tag array and antenna (tag-antenna distance) and the actual location of the entire setup, respectively. When the domain appears in the training set, RFree-GR achieves an average accuracy of 99%. When the domain is new, the accuracy can yield 89.03% , 90.21% and 88.38% for new user and new environment(new setup and new position), respectively. We compare RFree-GR with traditional machine learning methods and deep learning methods without adversarial learning. The result shows that our system outperforms the other methods by a large margin under domain-independent conditions.

Our contributions can be summarized as follows: (1) We exploit the multi-tag array to capture the spatio-temporal movements of gestures and Multimodal CNN to abstract multimodal features. This endows our system with the ability to sense complex and fine-grained gestures. (2) We exploit adversarial learning to remove domain-specific information in the received signals while retaining the gesture-specific information. (3) We implement RFree-GR on commercial RFID devices and

conduct extensive experiments. The results demonstrate the superior performance of RFree-GR for gesture recognition under domain-independent conditions.

The remainders of this paper are organized as follows. We first review the related works in Section 2. Then after introducing the preliminaries in Section 3, we illustrate the detailed design of the gesture system in Section 4. Furthermore, we present the implementation and evaluation in Section 5. Finally, we show the discussion of future work and the conclusion of this paper in Section 6 and Section 7, respectively.

2 RELATED WORKS

2.1 Gesture recognition

Existing gesture recognition can be mainly categorized into three classes: wearable-sensor based, computer-vision based and wireless based.

Wearable-sensor based methods leverage the sensors embedded in wearable devices to capture the hand or finger movements. For example, researchers integrate gyroscopes and accelerometers into a glove to track finger movements [15]. ArmTrak [23] fuse inertial signals of a smartwatch and the anatomy of arm joints to trace the geometric motion of the arm. Some other researchers leverage wristbands to enable fine-grained hand gesture recognition [13, 14] or sign language recognition [6, 42]. Although wearable devices are becoming more and more popular nowadays, people still feel inconvenient or forget to wear it for gesture recognition. Computer-vision based methods leverage cameras to capture images of gestures. The rapid development of deep learning substantially boosted the accuracy of these methods in activity [31] or gesture recognition [18]. Researchers also use depth-cameras like Kinect [44] or Leap Motion [3, 20] to further enhance the performance. The key limitation of vision based methods is their high sensitivity to lighting conditions as well as the privacy leakage issues.

Wireless based methods leverage ubiquitous wireless infrastructures to provide device-free and pervasive gesture recognition. Extensive researches have been made using WiFi [17, 28, 34], RFID [1, 2, 29, 43, 47], ultrasound [9, 33], and millimeter radar [32]. For example, E-eyes [34] recognizes both whole-body activities and coarse-grained gestures using channel state information of WiFi signals. SignFi [17] and WiFinger [25] also exploit WiFi signals to recognize frequently used sign gestures and fine-grained finger gestures, respectively. LLAP [33] uses speakers and microphones on mobile devices to perform device-free tracking of a hand/finger within millimeter accuracy. The major challenge of wireless based methods is their high dependence on users and environments [10, 45], which usually degrades their performance dramatically when it comes to a new domain, seriously limiting their acceptability in practical use.

2.2 RFID-based sensing

RFID has always been used for identification of objects, but recent research has shown that there is rich information implicated in the RF signals, which could be used for localization [35, 37], activity recognition [2, 38], human identification [7, 41] and vital sign monitoring [46]. For example, Tagoram [37] leverages COTS RFID tags and readers for object localization and tracking. RF-IDraw [30] track hand movements within high precision by attaching tags on fingers. To avoid the inconvenience of tag-on-body, much efforts have been made on tag-free sensing. TagFree [2] extracts the signal angle-of-arrival (AOA) information from a multi-antenna array to enable device-free activity recognition. RFree-ID [41] and Au-Id [7] leverage the tag array to capture the discriminative features among individuals for device-free user identification. ReActor [43] enables low-latency gesture recognition with machine learning algorithms. However, existing device-free RFID sensing solutions mainly focus on coarse-grained or simple gestures, rather than fine-grained

and complex gestures. Besides, these approaches also suffer seriously with domain-independent conditions.

2.3 Domain adversarial learning

Domain adversarial learning is to encourage a neural network to learn a robust representation which is discriminative to learning task on the source domain, but uninformative to interference domains. Based on the idea, Ganin et al. [4] first propose the basic domain adversarial training methods to tackle the domain adaptation problem. To further improve the domain adaptation performance, Zhao et al. [45] develop a conditional adversarial architecture which can remove the individual and condition specific information while retain the information relevant to the predictive task. Besides, Jiang et al. [10] propose a modified adversarial network to remove environments and subjects specific information for device free human activity recognition. These work treats all interference domains as a blend, handled with only one domain discriminator. Different from them, our work designs multiple domain discriminators, which brings better convergence speed and scalability.

3 PRELIMINARIES

In this section, we first present the characteristics of ASL, and then construct the RFID sensing model. Based on that, we discuss the sensing ability and domain diversity for gesture recognition.

3.1 American Sign Language

American sign language is expressed by complex gestures. Concretely speaking, these gestures consist of four components, i.e. handshape, palm orientation, location and movement [26]. Any change of these components may lead to a different sign, and thus many different sign gestures have similar arm, hand, or finger movements. For example, "uncle" and "aunt" have totally the same arm and hand movements of twisting the hand a couple of times, and the only difference exists in the handshape. Another example is "want" and "don't want", the only difference exists in their ending movements. In addition, some sign gestures involve small-scale finger movements. For example, "who" is to place the hand near the chin and bend the index finger twice. Therefore, to enable accurate and robust sign gesture recognition, we need to gather rich and fine-grained information in both temporal and spatial dimensions to comprehensively profile the gesture changes.

3.2 RFID Sensing Model

In a passive RFID system [1], when the reader transmit RF signals to interrogate tags, passive tags within range can harvest energy from the signals to make responses, and then the reader will receive the time-varying backscatter signal $S(t)$ which can be calculated as:

$$S(t) = a(t)e^{-j\theta(t)} = a(t)e^{-j(\theta_0 + \frac{4\pi d}{\lambda})} \tag{1}$$

where $a(t)$ and $\theta(t)$ are the amplitude and phase of the received signal, respectively. θ_0 is the initial offset, j is an imaginary unit and d is the length of the propagation path. As shown in Figure 2, when a user is performing gestures, some propagation paths remain invariant (static paths) while others change with human movements (dynamic paths). Thus $S(t)$ can be expressed as:

$$S(t) = S_s(t) + S_d(t) = a_s e^{-j\theta_s} + a_d(t)e^{-j\theta_d(t)} \tag{2}$$

Where S_s and S_d are static paths signals and dynamic paths signals, respectively. To get a more intuitive illustration on the signal variations, we depict the signal vectors on a complex plane, as

[1] https://www.impinj.com/platform/connectivity/speedway-r420

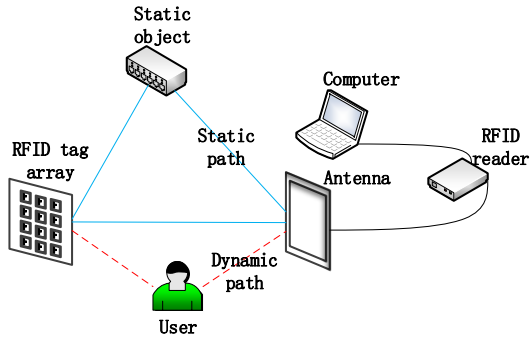


Fig. 2. Illustration of RFID signal propagation. The solid and dashed lines connecting the antenna and RFID tag array represent the static and dynamic paths, respectively.

shown in Figure 3. Within a short period time, S_s remain invariant, and the signal amplitude of S_d can also be considered as a constant but the phase changes dramatically. The phase change causes S_d to rotate with respect to S_s. For example, when the length of the dynamic path gets changed by λ, its phase will change (2π), which means S_d rotates 360 degrees. As such, the amplitude and phase of the resultant signal S, i.e. $|S|$ and $\angle S$, will also change with the changes of S_d. Present commercial RFID readers, such as Impinj Speedway in our implementation, can provide both the amplitude (RSS) and phase of the received signal, and we use both of them for gesture recognition.

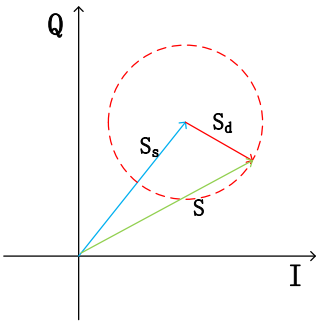
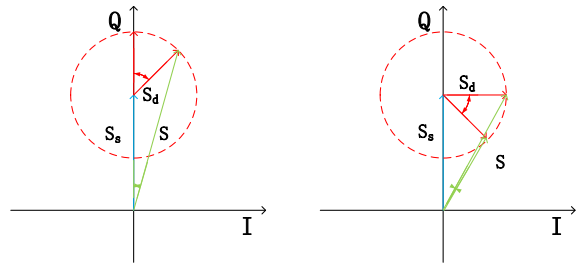


Fig. 3. The complex plane representation of signal.



(a) Good position for phase (b) Bad position for phase

Fig. 4. Illustration of blind zone.

3.3 The Sensing ability analysis for fine-grained gestures

Although the RF signal amplitude and phase are sensitive to subtle movements, we will show in this section that the sensitivity can be very weak at certain positions due to multipath fading effect. As shown in Figure 3, when the amplitude of S_d is constant, the signal variation of S are determined by the phase change of S_d. And when the change is more than 2π, the change of $|S|$ and $\angle S$ can also achieve the maximum which is $2|S_d|$ and $2arcsin\frac{|S_d|}{|S_s|}$, respectively. However, some sign

gestures involve subtle finger movements which may only introduce movements of less than $5cm$, and the corresponding phase change of S_d will be much less than 2π accordingly. In this case, the phase difference between S_s and S_d is the key factor which determines the sensing performance. As shown in Figure 4, for the same movement (the phase changes of S_d are same), when the phase difference of S_d and S_s is near 0 degrees (Figure 4a), the change of $\angle S$ is obvious, but when the phase difference of S_d and S_s is near 90 degrees (Figure 4b), the change of $\angle S$ is very slight. Therefore, the position of Figure 4a can be regarded as a good sensing position for phase, and the position of Figure 4b is a bad position which can be regarded as a blind zone for phase [40]. But for RSS, the sensing ability is the very opposite, which can also be concluded from Figure 4.

To further validate the above analysis on the sensing ability, we perform a preliminary experiment. A volunteer conducts the ASL sign gesture "who" for 5 times. As mentioned before, "who" is to bend the index finger twice. The RFID reader and tags are deployed on each side of the volunteer. In particular, the experimental setup looks like Figure 1. Instead of using a tag array, we use two tags and set the distance between tags to $4cm$. Figure 5 is the signal phase of these two tags. It is obvious that phase variations from Tag 1 is clear enough to identify the 10 repetitive small finger movements while that from Tag 2 is almost unresponsive. In conclusion, to sense the fine-grained gestures, we have to reduce sensing blind zone of RFID signals. On one hand, the sensing blind zone of the phase and the amplitude are different, so we combine phase and RSS together to enhance the sensing ability. On the other hand, we could set multiple RFID tags in different positions to form a RFID tag array for blind zone reduction. We will discuss the design of the tag array in Section 5.

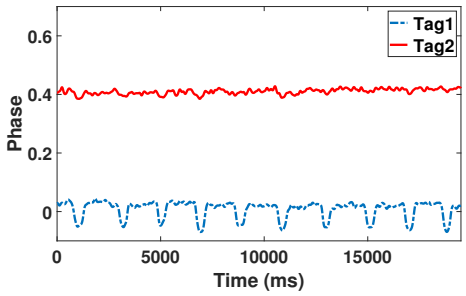


Fig. 5. Phase changes of two tags in different positions for the same gesture "who".

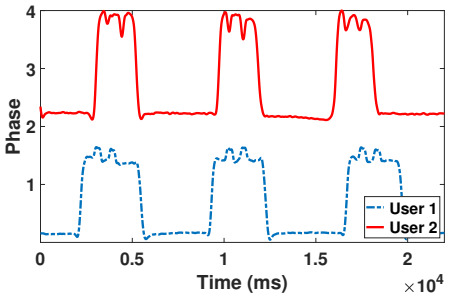


Fig. 6. Phase waveforms of two users performing the same gesture "uncle".

3.4 Domain diversity for gesture recognition

The received signal of the reader usually carry substantial information that is specific to the domains, i.e. users who conduct the gestures and environments where the gestures occur. On one hand, the signals, when being transmitted, may be reflected and diffracted by objects (e.g., wall, furniture) in the ambient environment. On the other hand, different human subjects with different ages, genders, body shapes and performing habits affect the signals in different ways. To illustrate this phenomenon, we asked two volunteers (denoted as User 1 and User 2) to perform "uncle" three times. The volunteers first raise their right hands and then perform the sign (extend the first two fingers and twist the hand a couple of times), after which they lower the hands to the initial position. Figure 6 shows the corresponding phase signal. Despite having some correlations, the signal fluctuation patterns in the sign performed stage differ considerably between two volunteers. The signal of User 1 exhibits two peaks when twisting the hand while the signal of User 2 exhibits

two valleys. The difference will become more significant when it comes to more complex gestures. Domain diversity extinguishes the application and popularization of many wireless sensing systems as it is impractical to collect enough training data for each domain.

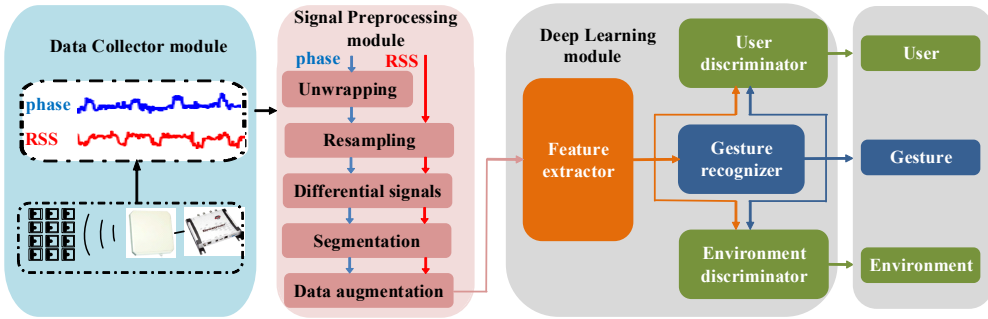


Fig. 7. System Overview.

4 RFREE-GR DESIGN

In this section, we first briefly introduce the overview of RFree-GR. Afterwards, we present all the core modules of RFree-GR in detail.

4.1 System Overview

As shown in Figure 7, RFree-GR consists of three major modules: data collector module, signal preprocessing module and deep learning module.

First, we leverage a RFID tag array to capture the raw phase and RSS data stream. Then, in addition to some necessary preprocessing, static interferences elimination and data augmentation are adopted to further improve the expression ability and diversity of data. After preprocessing, the data stream is fed into the deep learning module for gesture recognition. The deep learning module consists of feature extractor, user discriminator,environment discriminator and gesture recognizer. In particular, feature extractor is designed for abstracting intra-modality and cross-modality spatio-temporal representations. The feature extractor cooperates with the gesture recognizer to achieve high gesture recognition accuracy and simultaneously prevents the domain discriminators from distinguishing different domains. Meanwhile, the domain discriminators try to distinguish users and environments to play against the feature extractor. With adversarial learning, the feature extractor can learn domain-independent and gesture-discriminative representations and then the gesture recognizer can recognize gestures regardless of domain diversity.

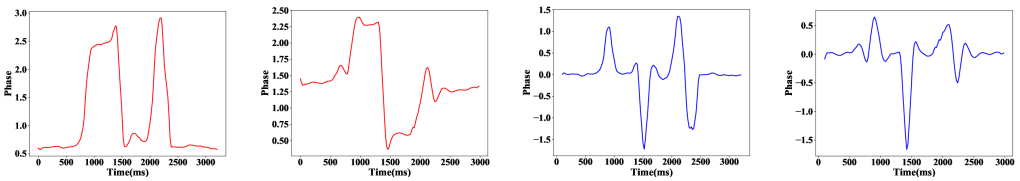
4.2 Signal Preprocessing

Due to various environment noise and working characteristics of RFID, it is unsuitable to directly feed the raw signals into the deep learning module. Therefore, we first pass the signals through the preprocessing module to enhance the sensing capability.

4.2.1 Phase unwrapping. The first step to process the raw phase measurements is phase unwrapping. The reason is that the signal phase reported by the reader is a periodic function ranging from 0 to 2π, termed as wrapped phase. In other words, the phase signal reported by the reader is not the true phase. The difference between the true phase signal and the phase signal reported by the reader is an integer multiple of 2π. Consequently, it is necessary to unwrap phase signal to remove the phase ambiguity. Specifically, we adopt the method in [47], which assumes that the absolute difference of two adjacent reading phase values is smaller than π, to unwrap the phase values. This is reasonable given that the frequency of human movements is well less than that of the reader interrogation. After the phase unwrapping, the weighted moving average filter, which is commonly used with time series data to eliminate short-term fluctuations and highlight long-term trends or cycles, is employed to reduce the high-frequency noise in phase and RSS.

4.2.2 Resampling. In RFID communications, RFID tags reply unevenly spaced in time domain due to RFID tags collision, packet loss or other delays, So the phase and RSS is not uniform in the time domain [5]. To deal with this, we adopt linear interpolation with 20ms between consecutive values to resample the data stream at a frequency of 50Hz.

4.2.3 Static interferences elimination. As described in section 3.2, the signals S received by the reader consist of static signals S_s and dynamic signals S_d. The dynamic signals S_d refer to the signals reflected from the dynamic gesture. And the static signals S_s denote signals that are reflected from static surfaces in the environment, such as non-moving objects or the user's body. The static signals are irrelevant or even harmful to gesture recognition, especially when transfering to a new domain. To eliminate these static interferences, we set the the differential signal $DS(t)$, which refers to the difference of signals between two consecutive samples (at time t and $t+1$), as the representation of gestures. In order to illustrate the effect of static interferences elimination, we asked a volunteer to perform "want" in two different environments (different surrounding reflectors). The result is shown in Figure 8. It can be observed that different environments will result in different waveforms in phase. But the differential phase still share the similar waveform in this case, proving the effect of static interferences elimination.



(a) Phase of "want" in environment 1 (b) Phase of "want" in environment 2 (c) Differential phase of "want" in environment 1 (d) Differential phase of "want" in environment 2

Fig. 8. The effect of static interferences elimination. Phase and differential phase waveforms of "want" performed by the same user in two different environments.

4.2.4 Segmentation. With the differential signal stream, we need to detect whether a gesture is occurring and segment it out. Since the duration and interval of gestures are various, a dynamic method is needed for the real-time detection and segmentation. As shown in Figure 8c, the differential signal fluctuates obviously when there is a gesture while remains stable when there is no gesture. In addition, the duration of the gesture is roughly between 1s and 3s. The two observations inspire us to use the variance to detect a gesture automatically. Given the differential

signals $DS(t)$, we set a small time window $[t - \frac{T}{2}, t + \frac{T}{2}]$ at time t with length T ($T = 0.5s$ in this paper). Then we can calculate the mean standard deviation of the multi tags at time t as $\sigma_t = \sqrt{\frac{1}{N \cdot M} \sum_{i=1}^{M} \sum_{j=t-\frac{T}{2}}^{t+\frac{T}{2}} (DS_i(j) - \bar{DS_i})^2}$, where M is the number of tag, N is the number of samples within this time window and $\bar{DS_i}$ denotes the average value of the these samples corresponding to tag i. In this way, we measured the mean standard deviation of all kinds of gestures and picked the smallest one (denoted as σ_{min}). To prevent the possibility of appearing a gesture with a smaller σ_t, we set the variance threshold to $\delta_\sigma = 0.7 \cdot \sigma_{min}$. Besides, in order to avoid the interference of accidental activities, the duration threshold δ_T (set as $0.8s$) is required in this paper. Then we can segment a gesture as following rules.

$$Maximize : t_{end} - t_{start} \tag{3}$$

$$\begin{aligned} Subject\ to : \sigma(t_i) \geq \delta_\sigma\ for\ \forall\ t_i \in [t_{start}, t_{end}] \\ t_{end} - t_{start} \geq \delta_T \end{aligned} \tag{4}$$

The $[t_{start}, t_{end}]$ part of the differential signal stream is detected as a gesture.

4.2.5 Data augmentation. Since deep neural network is data-hungry, we can apply a data augmentation scheme to enrich the diversity of data and enhance the robustness of RFree-GR. In this paper, we take into account the fact that different users have different gesture speeds. Specifically, we refer to the time warping technique in speech recognition [12], to design our data augmentation mechanism. Given a gesture profile $DS(t)$, we expend or contract it α times along the time axis. In this paper, we empirically set $\alpha \in \{0.6, 0.8, 1.2, 1.4\}$, thereby generating a new set profiles $S = \{DS(\alpha t) | \alpha \in \{0.6, 0.8, 1.2, 1.4\}$. In this way, we increases the dataset size by 5 times.

4.3 Deep Learning

After preprocessing, the data is fed into the deep learning module to enable feature extraction and gesture recognition. This module is the core part of RFree-GR consisting of feature extractor, two domain discriminators and gesture recognizer. In this subsection, we detail the design of this module.

4.3.1 Feature Extractor. The commercial RFID readers usually provide two signal indicators (i.e. RSS and phase) which are sensitive to gesture or other human activities. According to what we described in Section 3.3, the phase and the RSS information can complement each other to enhance the sensing performance, especially in blind zones. Existing methods for gesture recognition either only extract features from one modality, or directly concatenate the two modality streams and take them as input for the classification algorithms. However, as signal of each modality has different noise distributions, spatio-temporal representations and correlation structures, naive fusion is incline to miss the local interactions within each modality as well as the global interactions across modalities. With this in mind, we leverage a Multimodal Convolutional Neural Network to extract the high level representations from preprocessed data.

As shown in Figure 9, the feature extractor consists of two individual three-layer CNNs and a one-layer CNN. Specially, the two individual three-layer CNNs are used for exploiting the temporal and spatial information contained in phase and RSS, respectively. Besides, the final one-layer CNN is used to merge representations of phase and RSS, and extract across-modality information about gestures.

The preprocessed signal stream is a series of matrices with a size of $(t \times s)$, where t and s are the length of the time dimension and spatial feature dimension corresponding to each RFID tag, respectively. The two individual three-layer CNNs of phase and RSS modality share the same

network architecture but have individual optimized parameters. In particular, we adopt 2D CNN and the kernel size of each layer are 8×6, 8×3, and 5×3, respectively. We also apply pooling along the time dimension to reduce the size of the representation.

Through the individual CNNs, the high level representations of phase and RSS signals are extracted. Then they are merged to construct a new dimension (denoted as height). Therefore, the across-moduality relationships are implicated in the dimension of height. To fully exploit the across-modality information, we employ 3D kernel with shape $t \times s \times h$ (empirically set to $3 \times 3 \times 3$) in the merged convolution layer, where t, s, and h is the size of the kernel along time, spatial feature and height dimension, respectively. Finally, we flatten the output of the merged convolution layer for later processing.

For each convolution layer, RFree-GR leverages ReLU [19] as the activation function. Besides, we employ batch normalization (BN) [8] at each convolution layer to accelerate training and mitigate the over-fitting problem of the network. In general, we can get the high level representation R_M through feature extractor as follow:

$$R_M = FE(X, \theta_{fe}) \tag{5}$$

where X denotes the input signals and θ_{fe} is the set of all parameter in the feature extractor FE.

4.3.2 Gesture recognizer. Since the feature extractor has learned high level features, the gesture recognizer can simply use the fully-connected layer and softmax layer to achieve classification. In general, the gestures prediction probability vector G_P can be expressed as:

$$G_P = GR(R_M, \theta_{gr}) \tag{6}$$

where θ_{gr} is denoted as the set of all parameters in gesture recognizer GR. In addition, the cross entropy function can be used to calculate the loss L_g between the predictions G_P and the ground truths G_T as follows:

$$L_g(\theta_{fe}, \theta_{gr}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} G_P^{ij} \log(G_T^{ij}) \tag{7}$$

where N and J are the number of gesture samples and categories, respectively.

4.3.3 Domain discriminator. Similar to the gesture recognizer, we consider the tasks of these two domain discriminators as classification problems. In particular, the two domain discriminators take the concatenation of the high level features R_M and output of gesture recognizer G_P as input (denoted as R_{MG}) which can be expressed as follows:

$$R_{MG} = R_M \oplus G_P \tag{8}$$

where $\oplus$ means the operation of concatenation. Then the fully-connected layer and softmax layer are leveraged to map the input R_{MG} to the user predictions or environment predictions, just like the gesture recognizer network. In general, the user and environment prediction probability vectors U_P and E_P can be calculated as follows:

$$U_P = UDD(R_{MG}, \theta_{ud}) \tag{9}$$

$$E_P = EDD(R_{MG}, \theta_{ed}) \tag{10}$$

where θ_{ud} and θ_{ed} are all the parameters of the user domain discriminator UDD and the environment domain discriminator EDD, respectively. Besides, the loss functions L_u and L_e of the two domain discriminators can be expressed as follows:

$$L_u(\theta_{fe}, \theta_{ud}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} U_P^{ik} \log(U_T^{ik}) \tag{11}$$

$$L_e(\theta_{fe}, \theta_{ed}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} E_P^{im}\log(E_T^{im}) \tag{12}$$

where N, K, M means the number of gesture samples, users and environments, respectively. In addition, U_T and E_T are the ground truths of users and environments.

4.3.4 Domain adversarial learning. To deal with domain-independent conditions, we apply a domain adversarial learning which can discard extraneous features specific to users or environments, while remaining all features related with the recognition task.

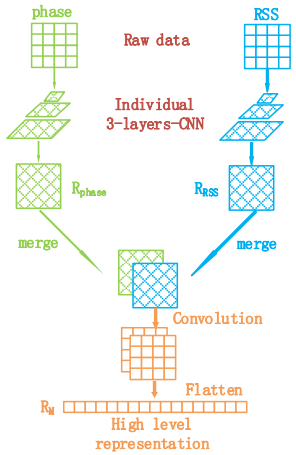


Fig. 9. Feature Extractor.

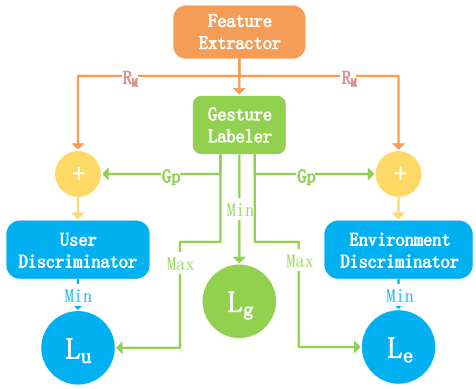


Fig. 10. Domain Adversarial Learning.

In particular, our domain adversarial learning regime involves four components: the feature extractor, the gesture recognizer and two domain discriminators, as shown in Figure 10. The feature extractor and gesture recognizer cooperate to predict gestures. However, if we only optimize the parameters θ_{fe} of the feature extractor and θ_{gr} of the gesture recognizer, for minimizing the loss L_g would not generalize well for new users and untrained environments. Therefore, we design a min-max game between the feature extractor and the discriminators. On the one hand, we take all the losses(L_g, L_u, L_e) into consideration when train the feature extractor and gesture recognizer network for predicting gestures, where the total loss L_t is calculated as follow:

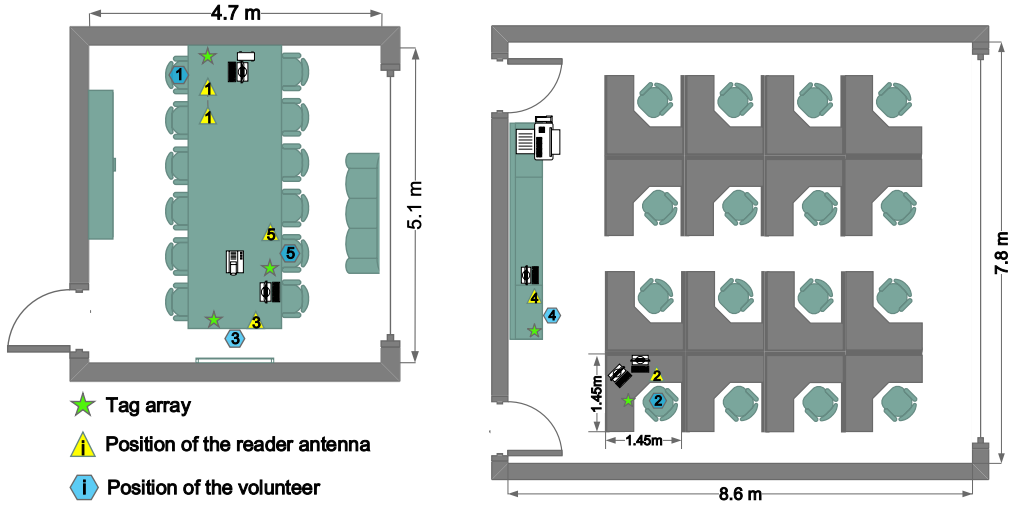
$$L_t(\theta_{fe}, \theta_{gr}, \theta_{ud}, \theta_{ed}) = L_g(\theta_{fe}, \theta_{gr}) - \alpha L_u(\theta_{fe}, \theta_{ud}) - \beta L_e(\theta_{fe}, \theta_{ed}) \tag{13}$$

The two coefficients α and β are predefined hyper-parameters which are used to control the impact of the user domain discriminator loss and the environment discriminator loss. Besides, the parameters θ_{ud} and θ_{ed} about domain discriminators are considered as constants in the process of training for predicting gestures. Therefore, the parameters θ_{fe} and θ_{gr} can be updated as follows:

$$(\theta_{fe}, \theta_{gr}) \leftarrow (\theta_{fe}, \theta_{gr}) - \lambda_1 \frac{\delta L_t(\theta_{fe}, \theta_{gr}, \theta_{ud}, \theta_{ed})}{\delta(\theta_{fe}, \theta_{gr})} \tag{14}$$

where λ_1 is the learning rate of the gesture optimizer.

On the other hand, we need to train the two domain discriminators in the same training period to update the parameters θ_{ud} and θ_{ed}. In the process of training for domain discriminators,



(a) Floor plan of meeting room(position 1,3,5) (b) Floor plan of office(position 2,4)

Fig. 11. Floor plan of environments scenario.

the parameters θ_{fe} are set to constants. Accordingly, the parameters θ_{ud} and θ_{ed} of the domain discriminators can be updated as follows:

$$(\theta_{ud}) \leftarrow (\theta_{ud}) - \lambda_2 \frac{\delta L_u(\theta_{fe}, \theta_{ud})}{\delta(\theta_{ud})} \tag{15}$$

$$(\theta_{ed}) \leftarrow (\theta_{ed}) - \lambda_3 \frac{\delta L_e(\theta_{fe}, \theta_{ed})}{\delta(\theta_{ed})} \tag{16}$$

where λ_2 and λ_3 are the learning rate of the two domain optimizers respectively. Specifically, for gesture optimizer, and two domain optimizers, we all adopt Adam optimizer [11] using default parameters (i.e., the learning rate 0.001, the exponential decay rate for the first-moment estimates of 0.9, the exponential decay rate for the second-moment estimates of 0.999 and the numerical stability parameter ϵ of 10^{-8}).

In conclusion, the process of training on the whole network is that the three optimizers update all parameters iteratively.

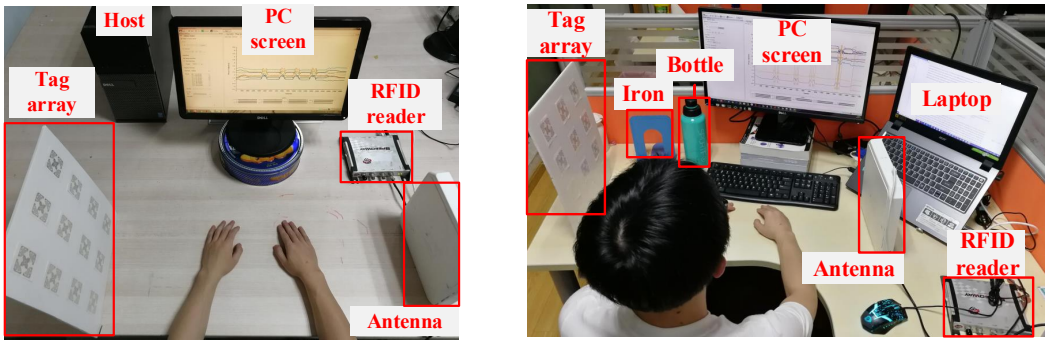
5 IMPLEMENTATION AND EVALUATION

In this section, we first introduce the experiment setup of RFree-GR. Then after introducing the dataset, we present micro-benchmark to evaluate the impact of system parameters and motivate the values selected in RFree-GR. And finally we present the detailed assessments of the various aspects of RFree-GR.

5.1 Experiment setup

RFree-GR is implemented with COTS UHF RFID devices. In particular, we leverage the Impinj Speedway R420 reader [2] equipped with a Laird S9028PCR antenna [3] to capture signals from the 4×3

[2]https://www.impinj.com/platform/connectivity/speedway-r420
[3]https://www.lairdconnect.com/part/s9028pcr



(a) Position 1 (b) Position 2

Fig. 12. Experiment Scenario.

Impinj H47UHF tag array (the size of tag array is discussed in section 5.3.1) with tags spaced 3cm apart. The reader operates at fixed frequency of $920MHz$. Since an antenna coverage is sufficient to cover the area used for sign language experiments, only one antenna is used. As shown in Figure 12, we place the tag array and antenna facing each other with a distance of 75 cm (the distance is discussed in section 5.3.2).

5.2 Dataset

In order to evaluate the performance of RFree-GR, we select 16 commonly used American Sign Language [26] words as the dataset which is a set of complex and fine-grained gestures. These signs nearly cover all the manual components in sign language, e.g. handshape, movement, palm orientation, location, one-hand and two-hand. Considering whether the gesture is one-handed or two-handed and whether it involves finger motion, the gesture set can be divided into 4 categories, as shown in table 1. Specifically, 15 volunteers (13 males and 2 females, age from 22 to 30, with different heights and weights) are invited to participate in the experiments. Before collecting data, they were asked to learn the 16 signs from ASL tutorial videos [26] until they could perform them correctly. Specifically, We collected 6080 samples which are organized into three parts: setup dataset, position dataset and user dataset. (1) The setup dataset is used to evaluate the robustness to the tag-antenna distance changes. We asked Volunteer 1 to conduct each gestures 10 times on five different setups in position 1 (Figure 12a). Then we selected a distance as the default setup on other experiments. (2) The position dataset is used to evaluate the robustness to the position changes of the entire setup. We asked Volunteer 1 to perform experiments 10 times in other four positions (position 2, 3, 4, 5). Combined with the data of volunteer 1 collected in position 1 with the default setup, the position dataset contained data collected form 5 different positions in total. (3) The user dataset is used to evaluate the robustness to user diversity, consisting of 4800 gesture samples where each of 16 gesture categories is performed 20 times by all 15 volunteers.

5.3 Micro-benchmark

5.3.1 Impact of tag array size. As we observed in Section 3.3, one or a few tags are not enough to capture comprehensive spatial information about gestures. Therefore, the size of tag array should be carefully considered, as there is a trade-off between performance and computational overhead. In

Table 1. The selected ASL words.

Category	Words
one hand with fingers motion	who, like, need
one hand without fingers motion	I, you, mother, uncle, aunt, food, drink
two hands with fingers motion	don't want, want
two hands without fingers motion	cold, more, time, clothes

order to cope with this problem, we conduct an experiment that performing the same set of gestures under 6 different tag array size ($1 \times 1, 2 \times 2, 3 \times 2, 3 \times 3, 4 \times 3, 4 \times 4$) and testing their performance separately. Figure 15 shows the average accuracy of within-user test and leave-one-user-out test under the six different tag array sizes. It is reasonable to assume that the accuracy would increase with a larger tag array, since more tags can capture richer spatial information related with gestures. In particular, when the size rises to 2×2, the accuracy of within-user test has already been very high, but the accuracy of leave-one-user-out test is still poor. Furthermore, when the size reach 4×3, the accuracy of both tests, are pretty good and there is almost no improvement as the size continues to improve. Therefore, we adopt 4×3 tag array in RFree-GR.

5.3.2 Impact of tag-antenna distance. As we descibed in Section 5.1, we place the tag array and antenna facing each other. The distance between tag array and antenna denotes tag-antenna distance which has impact on the phase and RSS of received signals. Considering the requirements of practical use and the characteristics of RFID, we select five kinds of distance ($70cm$, $75cm$, $80cm$, $85cm$, $90cm$) for evaluation. Specifically, we asked one volunteer to perform gestures under the five different tag-antenna distances. And the result is shown in Figure 17, the performance of tag-antenna distance of $75cm$ is the best, while the distance is too small or too large will lead to decreased accuracy. Therefore, we select $75cm$ as the tag-antenna distance in the following evaluation. In fact, the performance of system is acceptable when the tag-antenna distance change a little (eg. $70cm \sim 80cm$). And in the actual deployment, the requirement of tag-antenna distance is not particularly strict.

5.4 System performance

In this section, we evaluate the overall performance of RFree-GR, including within-user test, leave-one-user-out test, and new environment test.

5.4.1 Within-user test. Within-user test means that the data in the test set and the data in the training set belong to the same set of users. With 5-fold cross-validation, the average accuracy of within-user test reach over 99%. To better understand the classification errors, we present the confusion matrix of within-user test in figure 13. It is clear that all the signs can reach a very high accuracy with little errors. This is mainly because the data in the test set and the data in the training set share the same signal distribution. And it also indicates the reflected RF signals indeed carry the information of gesture which can be effectively extracted by RFree-GR.

5.4.2 Leave-one-user-out test. Leave-one-user-out test means taking a user's data as the test set and the remaining users' data as the training set, and then repeating this process for every distinct user. Figure 14 illustrates the results of leave-one-user-out test. We can observe that each one ranges from 82.29% to 96.15%, and the average accuracy of 15 volunteers is 89.03%. To gain a clearer view of the result, we show the average confusion matrix in figure 16. It can be observed that each gesture can be classified well. And the pairs of gestures that are often misclassified are those that are very similar. For example, "food" and "mom" both need to raise right hand and strike chin

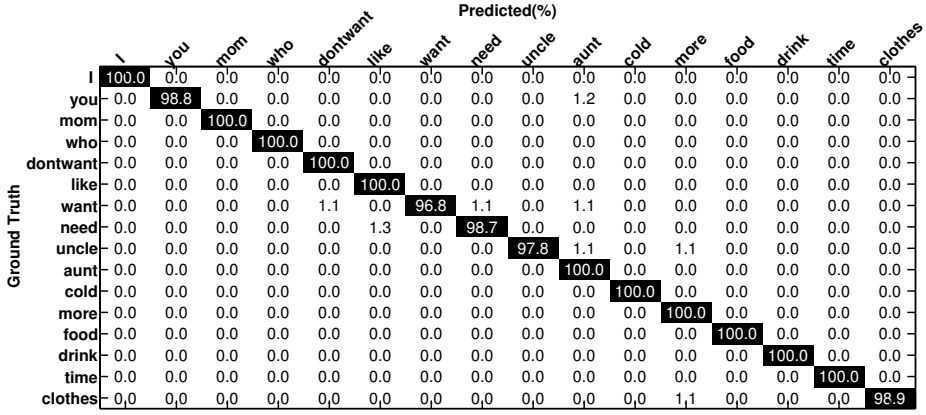


Fig. 13. Confusion matrix of within-user test.

twice while are only different in handshape. Overall, RFree-GR is able to accurately recognize each gesture in leave-one-user-out test.

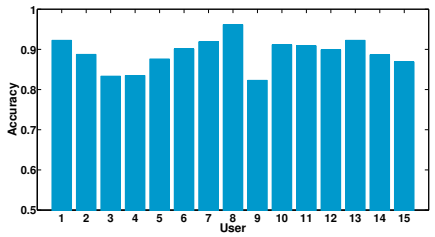


Fig. 14. Within-user Test.

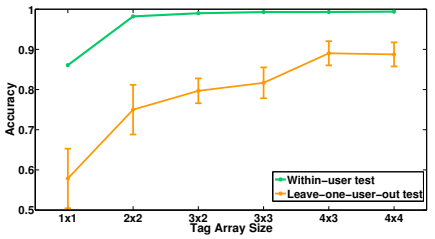


Fig. 15. Performance of different size of tag array.

5.4.3 New environment test. New environment test refers to the situation that the testing data is collected in a new environment. In this paper, different setups or different positions are considered as different environments. First, different setups refer to different distances between antenna and tag array (tag-antenna distance). As we described in section 5.3.2, we adopted leave-one-setup-out test on five different setups (with tag-antenna distance of 70cm, 75cm, 80cm, 85cm and 90cm). The result is shown in Figure 17. Specifically, the accuracy of 75cm is as high as 98.3%. Overall, RFree-GR can achieve an average accuracy of 90.21%, which indicates the robustness of RFree-GR across these disfferent setups. Second, we conduct another test (denoted as new posotion test) in five different positions in a meeting room and an office. The floor plans of the meeting room and the office are shown in Figure 11a and Figure 11b, respectively. And the locations of these five "positions" and their surroundings are also marked on the floor plans. The five positions are very different in the surrounding reflectors which will bring different multipath effects to the RFID signal. For example, in our experiments, position 2 (as shown in Figure 12b) in the office, is surrounded by many metal objects such as PC, laptop, bottle and iron book holder, while position 1 (as shown in Figure 12a) in the meeting room, is only with a PC screen and a host nearby. Then, a volunteer is asked to do

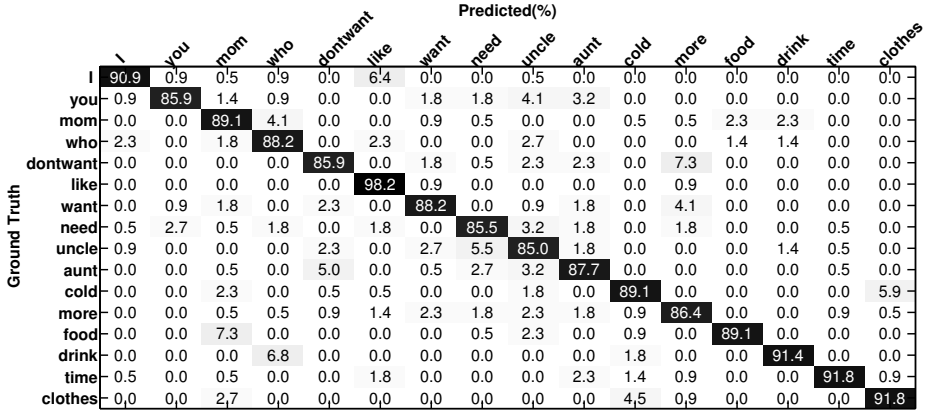


Fig. 16. Confusion matrix of leave-one-user-out test.

gestures in these five different positions for leave-one-position-out test (one position as the test set and the other four positions as the training set). As shown in Figure 18, although with a small amount of data (800 samples), the average accuracy is 88.38% (78% ∼ 98%), showing the strong robustness of our system to environmental changes. Furthermore, we consider a more practical and challenging situation where a new user in a new position with a new setup. Considering the dataset, there are 16 choices (1 user × 4 positions × 4 setups) for this test. As a result, RFree-GR can still achieve an average accuracy of 84.36%.

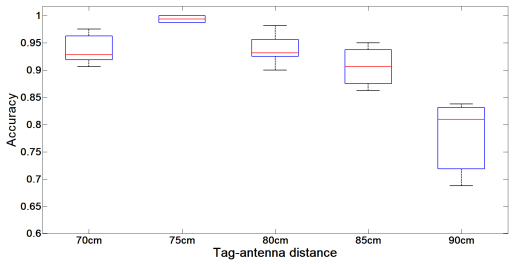


Fig. 17. New setup test.

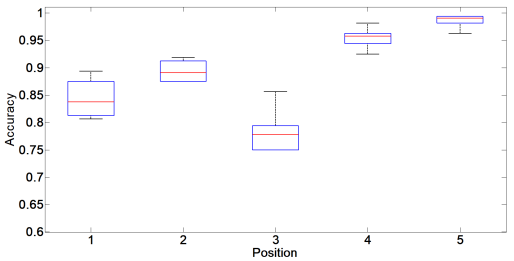


Fig. 18. New position test.

5.5 Significance of Signal Preprocessing

The signal preprocessing pipeline is significant to RFree-GR. We calculate the difference of signals at time axis to eliminate static interferences and adopt a data augmentation scheme to enhance the data diversity. To validate the significance of these two mechanisms, we set a series of experiments. In particular, we take three types of data(phase+RSS, differential signals, differential signals+data augmentation) as the input of the network, repectively, and perform leave-one-user-out test and new environment test(including new setup test and new position test). As table 2 shown, the differential signals can significantly improve system performance. And the combination of differential signal and data augmentation can further improve the performance of leave-one-user-out test. Since

the data augmentation mechanism in this paper is mainly set in consideration of the diversity of users but does not take into account the diversity of the environment, there is no improvement in the performance of the new environment test. In general, "differential signal+data augmentation" works best and is selected as the data preprocessing mechanism in this paper. Specially, it can improve the accuracy of 9.74%, 10.43% and 10.59% on leave-one-user-out test, new setup test and new position test.

Table 2. The performance under different signal preprocessing mechanisms.

Signal preprocessing mechanisms	Leave-one-user-out	New setup	New position
Phase+RSS	79.29%	79.78%	77.79%
Differential signals	85.27%	91.13%	89.24%
Differential signals+Data augmentation	89.03%	90.21%	88.38%

5.6 Significance of multimodal RFID Data Fusion

RFree-GR adopts Multimodal CNN to integrate phase and RSS streams reported by RFID reader. To evaluate the importance of multimodal RFID data fusion, we conduct a test. First, we replace the Multimodal CNN with a single three-layer CNN which has the same network parameters as multimodal case, and keep other network architectures unchanged. Then feed the modified system with phase-only data, RSS-only data, and the concatenation data of phase and RSS, respectively. Figure 19 illustrates the comparison results. It can be observed that using only a single modal signal (phase or RSS) is less accurate than the combination of them, especially for new setup test. This is due to the fact that both phase and RSS have their own sensing blind zone, which is explained in section 3.3. In addition, compared with the simple concatenation of phase and RSS, the multimodal RFID data fusion performs better, improving the accuracy of 2.37%, 2.46% and 1.88% on leave-one-user-out test, new setup test and new position test, respectively. This is because the simple concatenation of phase and RSS ignores the local interactions within each modality. These comparative results demonstrate the necessity of multimodal RFID data fusion in RFree-GR.

5.7 Significance of adversarial learning

In RFree-GR, we adopt adversarial learning to improve the performance of cross-domain test. In order to evaluate the effectiveness of adversarial learning in RFree-GR, we conduct the comparison with the variant of RFree-GR which does not engage in adversarial learning. And the results are shown in table 3. With the adversarial learning scheme, RFree-GR can futher improve the accuracy of the three tests by 1.77%, 6.46% and 2.13%, respectively. Considering that the signal preprocessing mechanism has dramatically improved the performance of cross-domain tests, the improvement brought by adversarial learning is quite obvious.

Table 3. The comparison of RFree-GR and its variant (without adversarial learning).

System scheme	Leave-one-user-out	New setup	New position
RFree-GR without adversarial learning	87.26%	83.75%	86.25%
RFree-GR	89.03%	90.21%	88.38%

5.8 Comparison with other classification algorithms

In RFree-GR, we adopt a hierarchical multimodal CNN with adversarial learning architecture to realize classification of gestures. To demonstrate the effectiveness of the classification algorithm, a series of comparative experiments are constructed. First, we consider some traditional classification algorithms, i.e., support vector machine (SVM), k-nearest neighbors (KNN). In addition, we adopt ReActor (Statistic Attributes+Wavelet coefficient->Random forest) [43], SignFi (CNN) [17] and TagFree (CNN+LSTM) [2] which are the state-of-the-art activity recognition systems based on wireless signal, to join the comparison. The comparison results are illustrated in Figure 20. Apparently, the performance of these traditional machine learning methods on these cross-domain tests is very poor. By contrast, the algorithms based on deep neural networks perform better due to the strong representation ability of deep learning. In particular, the CNN-based algorithms (SignFi, MCNN and RFree-GR) are better than CNN+LSTM (TagFree), let alone that CNN is more lightweight than LSTM (RFree-GR achieves up to 10X speedups in inference over the TagFree (CNN+LSTM)). RFree-GR yields the best performance in all the three tests (89.03%, 90.21% and 88.38% in leave-one-user-out test, new setup test and new position test), indicating the effectiveness of the classification algorithm used in RFree-GR.

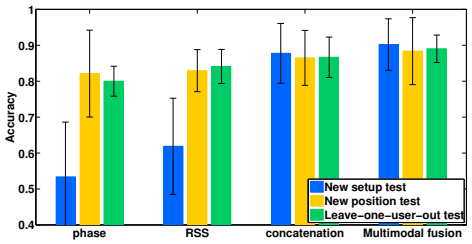


Fig. 19. Impact of multimodal RFID data fusion.

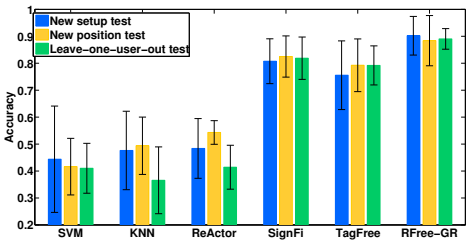


Fig. 20. Comparison of Different classifiers.

6 DISCUSSION

6.1 RFree-GR in the wild

In order to apply RFree-GR in real life, there are still some limitations. First, more quantities and more kinds of gesture data need to be collected to meet the requirements of specific applications. At the same time, the network may need to increase parameters to recognize more gestures.

Second, in an outdoor setting, the performance of RFree-GR will be influenced by the activities of people nearby. Since the signal is emitted from the front of the antenna at a radiation angle of 120 degrees, the coverage of the signal is on the front side of the antenna. For the region between the tag array and the antenna is relatively small ($75cm \times 50cm$), indicating little possibility of other people moving. For the region behind the tag array, we can attach the absorbing material on the back of the tag array to reduce possible interference.

Finally, RFree-GR requires a strict relative position between the user and the device. In the future, we plan to leverage the tag array and the antenna array to localize and track the hand movements of the user. Then we can use the trajectory of the hands to enable gesture recognition that is insensitive to user position.

6.2 Sequence recognition

Compared with isolate gesture recognition, gesture sequence recognition (e.g., a sign sentence) is more meaningful in practice. Some existing works shine a light on this issue. For example, some researchers leverage Long Short Term Memory (LSTM) with Connectionist Temporal Classification (CTC) [24] or encoder-decoder network with attention mechanism [27] to realize a speech recognition, which shows the potential to handle sequential gesture recognition. Therefore, it is possible to replace the gesture recognizer with a sequence recognizer based on LSTM+CTC or encoder-decoder network, to expand our system into sequence recognition.

6.3 Other gestures or other sensing mechanisms

RFree-GR is also suitable for other gestures. For example, we have conducted some experiments on the alphabet writing gestures. Specifically, five volunteers were asked to write 10 letters (from 'a' to 'j') 20 times in position 1 (Figure 12a). These writing gestures are performed by moving hand in mid-air while upper arm remained relatively stationary. And the moving range of these writing gestures is about 30cm. Through leave-one-user-out cross-validation, we are glad to find the average accuracy can reach 90%. The encouraging results indicate the feasibility of RFree-GR on the recognition of other gestures.

Moreover, RFree-GR has strong scalability as we can combine the data from other sensors like inertial sensors or cameras easily by its data fusion design. Besides, when a new interference domain appears in other scenario, we can just add another domain discriminator to cope with it.

7 CONCLUSION

In this paper, we propose RFree-GR, a device-free RFID-based system that enables domain-independent complex and fine-grained gesture recognition. We leverage a RFID tag array to comprehensively capture the spatio-temporal changes of gestures. After a series of signal preprocessing, we adopt feature extractor, gesture recognizer and two domain discriminators for adversarial learning to learn the domain-independent and gesture-discriminative features, which can achieve accurate and robust gesture recognition. Extensive experiments prove that RFree-GR can yield an average accuracy of 99% in domain-dependent conditions, and 89.03% , 90.21% and 88.38% for new user and new environment (new setup and new position), respectively. Given its innovative solution and promising performance, we believe RFree-GR has made a significant contribution to the advancement of gesture based Human-Computer Interaction.

ACKNOWLEDGMENTS

We sincerely thank all reviewers for their insightful suggestions. This work is supported by Shanghai RFID Engineering Technology Research Center.

REFERENCES

[1] Han Ding, Chen Qian, Jinsong Han, Ge Wang, Wei Xi, Kun Zhao, and Jizhong Zhao. 2017. Rfipad: Enabling cost-efficient and device-free in-air handwriting using passive tags. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 447–457.

[2] Xiaoyi Fan, Wei Gong, and Jiangchuan Liu. 2018. TagFree Activity Identification with RFIDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 7.

[3] Biyi Fang, Jillian Co, and Mi Zhang. 2017. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 5.

[4] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. 1180–1189.

[5] EPC Global. 2008. EPC radio-frequency identity protocols class-1 generation-2 UHF RFID protocol for communications at 860 MHz–960 MHz. *Version* 1, 0 (2008), 23.

[6] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. SignSpeaker: A Real-time, High-Precision SmartWatch-based Sign Language Translator. (2019).

[7] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-Id: Automatic User Identification and Authentication through the Motions Captured from Sequential Human Activities Using RFID. 3, 2, Article 48 (2019), 26 pages. https://doi.org/10.1145/3328919

[8] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[9] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 15.

[10] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 289–304.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association.*

[13] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 338.

[14] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 321–333.

[15] Kehuang Li, Zhengyu Zhou, and Chin-Hui Lee. 2016. Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. *ACM Transactions on Accessible Computing (TACCESS)* 8, 2 (2016), 7.

[16] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 142.

[17] 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 (2018), 23.

[18] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4207–4215.

[19] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

[20] Leigh Ellen Potter, Jake Araullo, and Lewis Carter. 2013. The leap motion controller: a view on sign language. In *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*. ACM, 175–178.

[21] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-wave Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[22] Stefano Savazzi, Stephan Sigg, Monica Nicoli, Vittorio Rampa, Sanaz Kianoush, and Umberto Spagnolini. 2016. Device-free radio vision for assisted living: Leveraging wireless channel quality information for human sensing. *IEEE Signal Processing Magazine* 33, 2 (2016), 45–58.

[23] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*. ACM, 85–96.

[24] Hagen Soltau, Hank Liao, and Hasim Sak. 2016. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. *arXiv preprint arXiv:1610.09975* (2016).

[25] Sheng Tan and Jie Yang. 2016. WiFinger: leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*. ACM, 201–210.

[26] American Sign Language University. [n.d.]. BASIC ASL. http://lifeprint.com/asl101/pages-layout/concepts.htm

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[28] Raghav H Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-user gesture recognition using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 401–413.

[29] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. 2018. Multi-touch in the air: Device-free finger tracking and gesture recognition via COTS RFID. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1691–1699.

[30] Jue Wang, Deepak Vasisht, and Dina Katabi. 2015. RF-IDraw: virtual touch screen in the air using RF signals. *ACM SIGCOMM Computer Communication Review* 44, 4 (2015), 235–246.
[31] Minsi Wang, Bingbing Ni, and Xiaokang Yang. 2017. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3048–3056.
[32] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 851–860.
[33] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 82–94.
[34] 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 the 20th annual international conference on Mobile computing and networking*. ACM, 617–628.
[35] Huatao Xu, Dong Wang, Run Zhao, and Qian Zhang. 2019. AdaRF: Adaptive RFID-Based Indoor Localization Using Deep Learning Enhanced Holography. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 113 (Sept. 2019), 22 pages. https://doi.org/10.1145/3351271
[36] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. 2019. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal* 6, 6 (2019), 10763–10772.
[37] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 237–248.
[38] Lei Yang, Qiongzheng Lin, Xiangyang Li, Tianci Liu, and Yunhao Liu. 2015. See Through Walls with COTS RFID System!. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*. Association for Computing Machinery, New York, NY, USA, 487–499. https://doi.org/10.1145/2789168.2790100
[39] Hui-Shyong Yeo, Juyoung Lee, Hyung-il Kim, Aakar Gupta, Andrea Bianchi, Daniel Vogel, Hideki Koike, Woontack Woo, and Aaron Quigley. 2019. WRIST: Watch-Ring Interaction and Sensing Technique for Wrist Gestures and Macro-Micro Pointing. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–15.
[40] Youwei Zeng, Dan Wu, Ruiyang Gao, Tao Gu, and Daqing Zhang. 2018. Fullbreathe: Full human respiration detection exploiting complementarity of csi phase and amplitude of wifi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–19.
[41] Q. Zhang, D. Li, R. Zhao, D. Wang, Y. Deng, and B. Chen. 2018. RFree-ID: An Unobtrusive Human Identification System Irrespective of Walking Cofactors Using COTS RFID. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–10.
[42] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2019. MyoSign: enabling end-to-end sign language recognition with wearables. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 650–660.
[43] Shigeng Zhang, Chengwei Yang, Xiaoyan Kui, Jianxin Wang, Xuan Liu, and Song Guo. 2019. ReActor: Real-time and Accurate Contactless Gesture Recognition with RFID. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[44] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia* 19, 2 (2012), 4–10.
[45] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. 4100–4109.
[46] R. Zhao, D. Wang, Q. Zhang, H. Chen, and A. Huang. 2018. CRH: A Contactless Respiration and Heartbeat Monitoring System with COTS RFID Tags. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. 1–9.
[47] Yongpan Zou, Jiang Xiao, Jinsong Han, Kaishun Wu, Yun Li, and Lionel M Ni. 2016. Grfid: A device-free rfid-based gesture recognition system. *IEEE Transactions on Mobile Computing* 16, 2 (2016), 381–393.

Received July 2020; revised August 2020; accepted September 2020