Robust WiFi-enabled Device-free Gesture Recognition via Unsupervised Adversarial Domain Adaptation

Han Zou‡, Jianfei Yang‡, Yuxun Zhou†, Lihua Xie† and Costas J. Spanos†
‡ Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
† School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
Email: {hanzou, yxzhou, spanos}@berkeley.edu, {yang0478, elhxie}@ntu.edu.sg

Abstract—Accurate human gesture recognition is becoming a cornerstone for myriad emerging applications in human-computer interaction. Existing gesture recognition systems either require dedicated extra infrastructure or users active cooperation. Although some WiFi-enabled gesture recognition systems have been proposed, they are vulnerable to environmental dynamics and rely on the tedious data re-labeling and expert knowledge each time being implemented in a new environment. In this paper, we propose a WiFi-enabled device-free adaptive gesture recognition scheme, WiADG, that is able to identify human gestures accurately and consistently under environmental dynamics via adversarial domain adaptation. Firstly, a novel OpenWrt-based IoT platform is developed, enabling the direct collection of Channel State Information (CSI) measurements from commercial IoT devices. After constructing an accurate source classifier with labeled source CSI data via the proposed convolutional neural network in the source domain (original environment), we design an unsupervised domain adaptation scheme to reduce the domain discrepancy between the source and the target domain (new environment) and thus improve the generalization performance of the source classifier. The domain-adversarial objective is to train a generator (target encoder) to map the unlabeled target data to a domain invariant latent feature space so that a domain discriminator cannot distinguish the domain labels of the data. In the phase of implementation, we utilize the trained target encoder to map the target CSI frame to the latent feature space and use the source classifier to identify various gestures performed by the user. We implement WiADG on commercial WiFi routers and conduct experiments in multiple indoor environments. The results validate that WiADG achieves 98% gesture recognition accuracy in the original environment. Furthermore, the proposed unsupervised adversarial domain adaptation is able to enhance the recognition accuracy of WiADG by 25% on average without the needs of labeled data collection and new classifier generation when implements it in new environments.

Index Terms—Internet of Things, gesture recognition, device-free, WiFi, adversarial domain adaptation.

I. INTRODUCTION

In the era of Internet of Things (IoT), gesture recognition is a critical underpinning to facilitate human-computer interaction in numerous smart home applications. For instance, home automation tasks, e.g. remote control of the household device and adjust the temperature and brightness level for personalized thermal comfort, can be achieved by gesture recognition. Furthermore, it can be implemented in entertain-
nition systems is how to improve the system portability and robustness over spatial and temporal dynamics. Since CSI describes how WiFi signal propagates from a TX to an RX through multiple paths, environment setup changing or operating in a new environment will modify the length of existing multi-paths or introduce new multi-paths. It will lead to drastic deviation of the real-time CSI readings from those data collected during offline training. Thus, the trained-once gesture classifier generated at particular time or location in one environment may not be able to serve as the reference for consistent gesture recognition in long-term deployments in itself or other environments. Existing CSI-based gesture recognition systems perform the entire training process (including training data collection labeling, and brand-new classifier generation) when the environmental scenario is altered. It is totally impractical because the process is time-consuming and labor-intensive to obtain enough label data to fine-tune a new classifier. Thus, an automatic and effective scheme that is able to neutralize the impact of environmental dynamics on gesture recognition performance is urgently desired.

To address the aforementioned issues, in this paper, we design a device-free WiFi-enabled adaptive gesture recognition system, WiADG, which is able to identify common human gestures with consistent high accuracy and robust to environmental dynamics via adversarial domain adaptation (ADA). Firstly, we develop an innovative CSI enabled IoT platform so that CSI readings can be obtained directly from COTS IoT devices instead of deploying laptops as RXs for CSI data acquisition. Moreover, instead of using CSI amplitudes, we leverage CSI phase differences across pairs of RX antennas to construct CSI frames as the input dataset for classifier generation. In the original environment (source domain), we design a convolutional neural network (CNN), that is able to extract the most discriminative local features from the CSI frames, to construct a source encoder (map CSI frames to a domain invariant latent feature space) and an accurate source gesture classifier. The training process does not require intensive human intervention because all the parameters in CNN are fine-tuned from end to end automatically.

We propose an unsupervised domain adaptation scheme inspired by Adversarial Discriminative Domain Adaptation (ADDA) [18] to tackle the environmental dynamics issue when WiADG operates in an untrained environment (target domain). Since our CSI enabled IoT platform is able to capture CSI frames in a non-intrusive manner with high sampling rate, we can easily obtain unlabeled data in the target domain. To minimize the domain discrepancy distance between source and target domain, we design a domain-adversarial objective function, which trains a generator (target encoder) to map the target data to the domain invariant latent feature space so that a domain discriminator cannot distinguish the domain labels of the data. After that, we utilize the trained target encoder to map the real-time target CSI frame to latent space and use source classifier to identify various gestures. We prototyped WiADG using COTS WiFi routers and conducted experiments in multiple real-world environments. Experimental results validate that WiADG can provide accurate gesture identification consistently against environmental dynamics without the tedious training process of data collection and labeling for new classifier generation in the target domain.

The rest of the paper is organized as follows. The related work is summarized in Section II. Section III introduces the detailed system design of WiADG. In Section IV, the experimental procedure is elaborated and the results are analyzed. We present our conclusions in Section V.

II. RELATED WORK

A. Domain Adaptation

The performance of conventional classifier will degrade severely when the data distribution of the training dataset (source domain) and testing dataset (target domain) are different. Domain adaptation aims to reduce the difference between the source and target feature distribution to improve generalization performance [19]. Recently, with the booming development of Generative Adversarial Network (GAN) [20], some researchers have proposed to construct an adversarial loss to accommodate the domain shift, which is commonly referred as adversarial domain adaptation (ADA) [21]. GAN consists of a generative model $G$ and a discriminative model $D$. The objective of $G$ is to synthesize images resembling real images, while the objective of $D$ is to distinguish real images from synthesized ones. $D$ and $G$ are trained jointly on the loss function in a min-max fashion using backpropagation, which update $D$ to maximize the likelihood of the discriminator being correct while update $G$ to generate realistic images to fool the discriminator. Similar to the learning configuration of GAN, the generator of ADA aims to fool the discriminator to make the target domain samples look like the source domain one, and the discriminator tries to identify the domain labels (source or target) instead of fake or real in GAN. Domain classifier is designed in [21] to predict the binary domain label of the inputs of the discriminator. ADDA [18] learns a discriminative representation using the labels in the source domain and then a separate encoding that maps the target data to the same space using an asymmetric mapping learned through a standard GAN loss without weights sharing.

III. SYSTEM DESIGN

A. Preliminary of CSI

WiFi signals propagate through multiple paths from TX to RX in indoor environments due to reflection, scattering and diffraction introduced by walls, doors, furniture, as well as the movements of occupants [22]. Different from the RSS which only captures the superimposition of multipath signals, CSI reveals fine-grained information about how the signal is propagated and interfered, including different time delays, amplitude attenuation, and phase shift of multiple paths on each subcarrier. Analyzing these signal propagation variations caused by human motions makes device-free gesture recognition feasible. In a nutshell, the signal can be modeled as a channel impulse response $h(\tau)$ and the OFDM receiver is able to provide a sampled version of the signal spectrum of
each subcarrier in the frequency domain, which contains both amplitude attenuation and phase shift as complex numbers. These measurements can be summarized as CSI:

\[ H_i = \| H_i \| e^{j\angle H_i} \]

where \( \| H_i \| \) and \( \angle H_i \) denote the amplitude and the phase of the CSI at the \( i^{th} \) subcarrier respectively.

**B. CSI enabled IoT Platform**

Most of existing CSI-based sensing systems adopt the Intel 5300 NIC tool [23] to extract the CSI data from laptops with external WiFi NIC cards. Requiring laptops as receivers severely limit them from large-scale implementation. To overcome this bottleneck, we develop a CSI enabled IoT platform so that the CSI measurements from regular data frames transmitted in the existing traffic can be obtained directly from the COTS IoT devices, such as commodity WiFi routers. OpenWrt is chosen as the OS for our platform since it is a lightweight and widely used Linux OS for embedded devices. We upgrade the Atheros CSI Tool [11] and develop a new OpenWrt firmware for IoT device for CSI acquisition. In addition, our platform reports CSI data on all the 114 subcarriers for 40 MHz bandwidth on 5 GHz central frequency, which provides much more information than conventional CSI tools. At each time instance, each TX-RX pair is able to provide \( N_{TX} \times N_{RX} \times 114 \) CSI amplitude and phase measurements, where \( N_{TX} \) and \( N_{RX} \) represent the number of TX and RX antennas.

**C. CSI Frames for Gesture Recognition**

By leveraging the designed CSI enabled IoT platform as introduced in Section III-A, we conducted a preliminary experiment by using 2 TP-LINK N750 wireless routers (one as TX and another one as RX) to evaluate whether distinct CSI measurements can be revealed for human gesture identification. The 2 routers were put 1m away on a table in a conference room. One volunteer performed six gestures, moving right and left, pushing and pulling, rolling right and left, near the line-of-sight of the TX-RX pair.

According to the experimental results, we observed that the CSI phase differences across pairs of RX antennas are more sensitive than CSI amplitude. As an example, Fig. 1 demonstrates CSI phase difference readings across one antenna pair and one subcarrier for 6 gestures, respectively. It can be clearly seen from Fig. 1 that when the volunteer was performing the gestures, disparate gestures perturbed in a distinct manner on the CSI phase difference readings. Pairs of relative gestures (moving right and left, rolling right and left, push and pull) display symmetrical patterns. Therefore, these observations verify that the CSI time-series data can be leveraged to identify various gestures. Moreover, from another perspective, the time-series phase difference measurements over multiple subcarriers can be treated as ‘video monitoring’ for gesture recognition. As highlighted with red lines in Fig. 1, the CSI time series data can be divided into small chunks with a window size \( \Delta t \). The data in each window form a CSI frame, that contains \( n \times m \) CSI pixels (\( n \) is the number of consecutive samples and \( m \) represents the number of distinct measurements of CSI phase difference readings). These CSI frames are served as input dataset for our designed gesture classifier of WiADG.

**D. Robust Gesture Recognition via Unsupervised Adversarial Domain Adaptation**

As introduced in Section II, several limitations prevent existing CSI-based gesture recognition systems [17], [16] for actual ubiquitous implementation. Firstly, labeled data are required to be collected whenever these systems are deployed in an untrained environment. After the tedious data collection process, expert knowledge is needed to design dedicated filtering technique and feature extraction method that work well in that environment for training dataset construction. Then, existing methods employ KNN and SVM [24], [25] to build up a brand-new classifier for gesture recognition. Their performance is vulnerable to temporal and environmental dynamics since the conventional machine learning techniques fail to update the classifier adaptively.

In this work, we design a device-free WiFi-enabled adaptive gesture recognition system, WiADG, that aims to realize accurate and robust human gesture identification consistently under spatial and temporal variations via adversarial domain adaptation (ADA). Inspired by the work presented in [18], we design the WiADG methodology consisting of three main steps, which are summarized in Fig. 2.

![Fig. 1. CSI phase difference readings across antenna pair RX\(_2\) and RX\(_3\) on subcarrier 11 for 6 common human gestures.](image-url)
The methodology of WiADG. Step 1: in the original environment (source domain), a source encoder and a source classifier are generated with the labeled source CSI frames. Step 2: a target encoder is trained through unsupervised adversarial domain adaptation to map the target CSI frames to the shared latent feature space such that a domain discriminator cannot distinguish the domain labels of the data. Step 3: the trained target encoder maps the target CSI frames to the domain invariant latent feature space and the source classifier recognize gestures during the implementation. The network parameters in solid line boxes are fixed and those in dashed line boxes are required to train.

**Step 1:** Suppose we collect $L$ CSI frames $X_s$ with labels $Y_s$ (the ground truth of gesture type) in an environment (referred to as the original environment, source domain). The first step of WiADG is to train a source representation mapping (source encoder), $M_s$, and an accurate source gesture classifier $C_s$. The objective can be summarized as the following optimization:

$$
\min_{M_s, C_s} \mathcal{L}_{C_s}(X_s, Y_s) = - \mathbb{E}_{(x_s, y_s) \sim (X_s, Y_s)} \sum_{l=1}^{L} [I[y_s = y_l] \log D(M_s(x_s))] 
$$

(1)

Fig. 3 illustrates the convolutional neural network (CNN) architecture we designed for the source encoder and source classifier. The goal of $M_s$ aims to extract discriminative local features from CSI frames and map them to a latent feature space. It is composed of a cascade of two pairs of convolutional layer and subsampling layer, followed by three fully-connected layers. The objective of convolutional layer is to exploit the local dependency features from input data. It extracts local features by using a bank of filters sliding over the input followed by nonlinear activation functions. We use rectified linear unit (ReLU) as the activation function in this work. The subsampling layer aims to reduce the dimensionality of the data while guaranteeing the invariance of feature maps by max pooling. The source gesture classifier $C_s$ consists of three fully connected layers, which is followed by the source encoder $M_s$.

The detailed dimension of each layer is illustrated in Fig. 3. To train the source encoder and the source gesture classifier, the network outputs are calculated forwardly and the cross-entropy loss between the predicted outputs and true targets are computed. Then, we utilize ADAM [26] as the optimizer to back-propagate the gradient layer by layer to update the parameters (weights and biases) in CNN. In this manner, the constructed $C_s$ is able to identify various gestures accurately in the original environment.

**Step 2:** This is the vital step of WiADG, which aims to recognize gestures in a brand-new environment (target domain) without the extra task of re-calibration and collection of labeled data in that environment. Since our CSI platform is able to collect data in a non-intrusive manner with high sampling rate, we can easily obtain unlabeled CSI frames in the target domain while user is performing gestures. These unlabeled CSI frames from the new environment is denoted by $X_t$. With these data, the following objective is to minimize the distance between the source and target mapping distributions $M_s(X_s)$ and $M_t(X_t)$ so that the source gesture classifier $C_s$ can be directly applied to identify various gestures in the new environment without the need to learn a separate target classifier $C_t$. We perform adversarial adaptation by learning a target representation mapping (target encoder) $M_t$ such that a discriminator $D$ cannot distinguish the domain label of encoded source and target samples. It is similar to the original GAN, that aims to generate fake image that is indistinguishable from the real image. In our case, the labels for the discriminator $D$ are domain labels (source and target) instead of fake and real. The adversarial loss can be formulated as follows,

$$
\min_D \mathcal{L}_D(X_s, X_t, M_s, M_t) = - \mathbb{E}_{X_s \sim X_s} [\log D(M_s(x_s))] - \mathbb{E}_{X_t \sim X_t} [\log (1 - D(M_t(x_t)))]
$$

(2)

The inverted label GAN loss [20] is employed to train the target encoder $M_t$ as follows,

$$
\min_{M_t} \mathcal{L}_{M_t}(X_s, X_t, D) = -\mathbb{E}_{X_t \sim X_t} [\log D(M_t(x_t))]
$$

(3)

It provides stronger gradients to the target mapping. In order to train the target encoder $M_t$ more effectively, we leverage the parameters of the source encoder $M_s$ learned in the Step 1 as an initialization for $M_t$ and fix $M_s$ during this adversarial learning process. In our design, the discriminator $D$ consists of 3 fully connected layers: 1024 hidden units - 2048 hidden units - binary label output. ReLU is employed as the activation function in these layers. The parameters in $M_t$ and $D$ are tuned jointly using backpropagation.

**Step 3:** During the implementation phase, we map the real-time CSI frames to the shared feature space through the target
encoder $M_t$ constructed in Step 2 firstly, and then adopt the pre-trained source gesture classifier $C_s$ to identify the gesture in the new environment (target domain).

To sum up, as described in Fig. 2, the first step of WiADG is to train a source encoder $M_s$ and a source classifier $C_s$ with the labeled source data by optimizing $L_{C_s}$ as described in equation 1. After that, we fix $M_s$ and learn a target encoder $M_t$ through adversarial learning, which aims to optimize $L_D$ (equation 2) and $L_{M_t}$ (equation 3) without revisiting the first objective equation. 1. In the phase of implementation, we utilize the trained target encoder $M_t$ to map the target CSI frame to the latent feature space and directly use the source classifier $C_s$ to identify various gestures.

IV. IMPLEMENTATION AND EVALUATION

A. System Setup

We prototyped WiADG using 2 TPLINK N750 routers (one serving as TX and the other as RX) and evaluated it in real indoor environments. The firmware of the routers was upgraded to our CSI enabled IoT platform so that the CSI measurements from regular data frames are reported directly from the RX. TX was operated on 5 GHz with 40 MHz channel bandwidth, which have higher opportunity to capture the detailed small-scale fading effects caused by subtle gestures, than 2.4 GHz with 20 MHz channel bandwidth. After receiving the data frames from TX, RX analyzes the data packet, extracts the CSI data and forwards them to a back-end computation unit through UDP. The computation unit using our experiments is a Think-pad laptop with Intel i7-4810MQ 2.80GHz CPU and 16GB RAM. It processes the CSI time-series data in real-time using Python. The sampling rate was 100 packets/s and linear interpolation was adopted to ensure the stationary interval of consecutive CSI values when there was a packet loss. We leveraged the CSI phase difference across 1 pairs of 3 antennas on RX router to construct the CSI frame. The dimension of each CSI frame is $400 \times 114$.

B. Data Collection

Experiments were conducted in 2 typical indoor environments (i.e. a conference room ($7m \times 5m$) and an office zone ($4.5m \times 5.6m$) as shown in Fig. 4) to validate the performance of WiADG under both same and disparate environment scenarios. TX and RX routers were put 1 m apart on a table as shown in Fig. 5 and 2 volunteers performed 6 common gestures, i.e. moving one hand right and left, rolling right and left, and push and pull. For each gesture at each testbed, 100 samples were collected on one day to train the source gesture classifier $C_s$ and other 100 samples were obtained on different days to reflect the temporal dynamics for testing purpose. In total, we acquired more than 2500 CSI frames to validate the gesture recognition accuracy and the effectiveness of WiADG over environmental dynamics.

C. Performance of WiADG in the Original Environment

Firstly, we evaluate the performance of WiADG in the original environment. In general, WiADG achieves an average
cross-validation gesture recognition accuracy of 98.3% and 98% in the conference room and the office zone, respectively. Its accuracy confusion matrices in these two environments are demonstrated in Fig. 6. Its performance is better in the conference room than the office zone because the environmental scenario is more complicated in the office zone. However, as shown in Fig. 6, the gesture identification accuracy of each category is at least 94%, validated the high stability of WiADG.

![Confusion matrices of gesture recognition accuracy using WiADG in the original environments (source domain).](image)

Fig. 6. Confusion matrices of gesture recognition accuracy using WiADG in the original environments (source domain).

We further compared its performance with 2 state-of-the-art CSI-based gesture recognition system, WiG [16] and WiAG [17]. WiG adopted wavelet denoising process to sanitize the raw data and constructed an SVM classifier to distinguish gestures. WiAG utilized principal component analysis (PCA) to denoise the raw CSI amplitude data, discrete wavelet transform to generate features and KNN to train the gesture classifier. Fig. 7 evaluates the true positive rate (TPR) of the three methods. TPR indicates the ratio of the number of times for correctly recognizing a gesture to the total number of gestures performed. As presented in Fig. 6, WiADG achieves the best accuracy for every gesture in both environments among the three approaches. It enhances the overall TPR over WiG and WiAG by 9.5% and 7.7% in the conference room, and 10.9% and 8.6% in the office zone, respectively. Since both WiAG and WiG are employed conventional machine learning approaches to build up the classifier, they cannot precisely extract the features that are related to gestures. On the other hand, the source encoder of WiADG, which is designed based on conventional neutral network, is able to explore the local dependency among the CSI frames for better representation.

**D. Gesture Recognition Performance of WiADG under Environmental Dynamics**

We further evaluate WiADG in a more realistic scenario where the testing environment (target domain) is different from the one (source domain) during the training stage. Table I compares the true positive rate (TPR) of WiADG and other methods under this circumstance. As demonstrated in Table I, the performance of all the three methods degraded severely if the source classifier of each method is directly applied in the new environment. It is explainable because multi-paths conditions of the original and the new environment are notably different, which leads to huge deviation of the CSI readings. Although the recognition accuracy of WiADG (source only) is slightly higher than WiG and WiAG, it is still too low for practical implementation.

As mentioned in the previous section, constructing a new classifier through data collection labeling and training is extremely time-consuming and labor-intensive. Thus, we performed the unsupervised adversarial domain adaptation as described in Section III-D for WiADG to accommodate the environmental dynamics. We learned a target encoder $M_t$ through adversarial learning to map the target CSI frames to the source latent space to minimize the impact of domain shift, and then leveraged the source classifier $C_s$ for gesture recognition. As shown in Table I, the proposed WiADG with domain adaptation (the last row in Table I) outperforms the existing methods (i.e. WiG and WiAG) by at least 30.5%. To justify the net contribution of unsupervised adversarial domain adaptation, we also compare it with the source only WiADG, i.e., the source encoder and source classifier obtained at step 1 as shown in Fig. 2. Comparing the last two rows of Table I, we observe that the major improvement is the consequence of the proposed novel method (step 2 and step 3 of WiADG). It provides compelling evidence to verify that the
proposed unsupervised adversarial domain adaptation method empowered the robustness of WiADG over environmental dynamics.

<table>
<thead>
<tr>
<th>TPR(%) Approach</th>
<th>Conf room → Office</th>
<th>Office → Conf room</th>
</tr>
</thead>
<tbody>
<tr>
<td>WiAG (Source Only)</td>
<td>49.8</td>
<td>36.1</td>
</tr>
<tr>
<td>WiADG (Source Only)</td>
<td>49.7</td>
<td>35.4</td>
</tr>
<tr>
<td>WiADG (Domain Adaptation)</td>
<td>50.7</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td><strong>83.3</strong></td>
<td><strong>66.6</strong></td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, we proposed WiADG, a device-free WiFi-enabled gesture recognition scheme that is able to identify human gestures accurately and consistently under environmental dynamics via unsupervised adversarial domain adaptation using only commodity IoT devices. To avoid the use of dedicated infrastructure, we developed a novel OpenWrt-based IoT platform to obtain fine-grained CSI time series data from IoT devices. After an accurate classifier is built up in the original environment (source domain), we designed an unsupervised adversarial domain adaptation scheme to reduce the domain discrepancy between the source and the target domain (new environment) and thus alleviated the influence of environmental dynamics. We trained a generator (target encoder) to map the unlabeled target data to a shared latent feature space so that a domain discriminator cannot distinguish the domain labels of the data through a domain-adversarial loss. To implement WiADG in the new environment, the trained target encoder was employed to map CSI frame to the latent feature space so that the source classifier can be used to identify gestures in the target domain. Real-world experiments with COTS WiFi routers were conducted and demonstrated that WiADG improves the robustness of the system under environmental dynamics without tedious labeled data collection and complicated new classifier construction process in new environments.

ACKNOWLEDGMENT

This research is funded by the Republic of Singapore National Research Foundation (NRF) through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore. The research is also partially supported by a 2018 Seed Fund Award from CITRIS and the Banato Institute at the University of California.

REFERENCES


