# A Robust Sign Language Recognition System with Sparsely Labeled Instances Using Wi-Fi Signals

Jiacheng Shang and Jie Wu
*Center for Networked Computing*
*Temple University, Philadelphia, USA*
*jiacheng.shang@temple.edu, jiewu@temple.edu*

***Abstract*—Sign language is important since it permits insight into the deaf culture and allows more opportunities to communicate with those who are deaf or hard of hearing. In this paper, we show that Wi-Fi signals can be used to recognize sign language with sparsely labeled training dataset. The key intuition is that sign language introduces different multi-path distortions in Wi-Fi signals and generates different unique patterns in the time-series of Channel State Information (CSI) values. Based on these observations, we propose a sign language recognition system called WiSign. Different from existing Wi-Fi signal-based human activity recognition systems, WiSign only requires a sparsely labeled training dataset. Two solutions based on transfer learning and semi-supervised learning are proposed to reduce the number of required labeled instances. We implemented WiSign using a TP-Link TL-WR1043ND Wi-Fi router and a Lenovo X100e laptop. The evaluation results show that WiSign can achieve a mean prediction accuracy of $87.01\%$ and $87.38\%$ for the transfer learning-based approach and semi-supervised learning-based approach, respectively.**

***Keywords*-human recognition systems, machine learning, signal processing.**

## I. INTRODUCTION

Sign language is important since it lends us insight into the deaf culture and bestows more opportunities to communicate with those who are deaf or hard of hearing. Since sign language mainly uses manual communication to convey meaning, we can use Human Activity Recognition (HAR) techniques to recognize them. Various systems have been designed for HAR using different devices and techniques, like camera, low-cost radar, and wearable sensors. However, they all have some limitations when applied in practice. For example, camera-based approaches need a line-of-sight and sufficient lighting. Camera-based approaches may also breach human privacy in some scenarios (e.g. in the bathroom). The operation range of the low-cost radar is limited to just tens of centimeters, which limits its deployment in a large room. Wearable sensor-based approaches can achieve a high accuracy with a low cost, but they ask the users to wear their sensor during the recognition, which is inconvenient and not practical in some applications (e.g. rescue scenarios).

In the past few years, researchers find that Wi-Fi signals can be leveraged to recognize various human activities. The key intuition is that different human activities will introduce different multi-path distortions in Wi-Fi signals and generate different patterns in the time-series of CSI values. With modified driver, we can collect these CSI estimations from the hardware. Also, thanks to the high data rate supported by modern commercial Wi-Fi devices, we can capture enough CSI measurements during each human activity. Based on this observation, researchers have proposed various systemsto recognize different human activities. For instance, CARM [1], proposed by Wang et al. can recognize different human activities based on spectrum analysis. Li et al. proposed WiFinger [2], which can achieve number text input in Wi-Fi devices by recognizing finger-grained gestures. Wang et al. proposed WiHear [3], which can trace mouth movements and recognize different words that people say. These systems follow the general structure of machine learning-based systems and generally have four stages: data collection, noise removal, feature extraction, and classification.

However, all existing Wi-Fi based approaches have some limitations. Some systems, like WiWho [4], could recognize different people with a high accuracy, but they are based on the generative model, like the decision tree. Such models have the potential requirement that the label distributions in the training dataset and the testing dataset should be the same. In real human activity recognition applications, this requirement is usually hard to satisfy. Generative models also tend to produce a significant number of false positives, which is particularly true for activities that are similar. The discriminative model is more practical in these cases since it enables the construction of flexible decision boundaries and does not require the same label distribution, which results in classification performances often superior to those obtained by purely probabilistic or generative models [5–7]. Several Wi-Fi based human activity recognition systems, like WiFall [8], have adopted the SVM model which is a discriminative model, but all of these systems support just two labels (e.g. falling down or not). In a real human activity recognition application, at least 3 activities should be supported. Moreover, most of the current machine learning-based systems require a large amount of labeled training dataset that is usually hard and expensive to get in practice.

In this paper, we propose a sign language recognition system using Wi-Fi signals called WiSign. WiSign can support basic sign language recognition like “Yes” and “Good Bye”.


Figure 1. Sign language [9]

Different from existing Wi-Fi signal-based human activity recognition systems, WiSign only requires a small labeled training dataset and avoids the potential risk of adopting the generative model. WiSign consists of two commercial Wi-Fi devices. One of them is used as a transmitter that keeps emitting signals and the other one is used to keep receiving those signals. When a user performs a specific hand or arm movements within the range of WiSign, WiSign recognizes the meaning based on the analysis of the variety of CSI waveforms. Two solutions based on transfer learning and semi-supervised learning are designed to reduce the number of required labeled instances. We consider two cases in which the new user can only provide a small labeled dataset. If the user can provide enough unlabeled training datasets, then we can use the idea from semi-supervised learning to leverage the knowledge in labeled instances to label those unlabeled samples; If the new user cannot provide extra unlabeled dataset, we can use the idea of transfer learning, which transfers the knowledge of other users to train a classifier for a new user. Based on the experimental results, our system can achieve similar even better sign language accuracy with sparsely labeled dataset. Besides, since our system leverages the knowledge from unlabeled data and others' similar datasets, the new classifier is still robust enough for the new user.

The key contributions of our proposed solutions can be summarized as follows:

- *Our semi-supervised learning-based solution makes full use of unlabeled data to improve the performance of learning.* Unlabeled instances are easy to obtain since they do not require humans' annotation efforts. In our system, a semi-supervised learning framework is adopted to label those unlabeled instances using the knowledge from labeled instances.
- *Our transfer learning-based solution makes full use of auxiliary data collected from others.* When training instances are very scarce, supervised learning is difficult. Besides, auxiliary instances (e.g. others' training datasets) are often available in our application. In our system, we calculate the similarities between two labeled instances and choose auxiliary data which are similar to the new user's training data.

The remainder of this paper is organized as follows: In Section II, we will introduce some existing wireless signal-based human activity recognition systems which use special hardware, Received Signal Strength (RSS), or CSI. In Section III, we will discuss the challenges we faced and the structure of WiSign. Signal preprocessing, feature extraction, and classification algorithms will be discussed in Sections IV, V, and VI. In Section VII, we will introduce our experiment implementation and analyze the evaluation results. The final conclusion and future work will be given in Section VIII.

## II. Related Work

### A. Wi-Fi based human activity recognition system

Existing wireless signal-based human recognition systems can be divided into 3 categories: Special hardware-based, RSS-based, and CSI-based.

*1) Special hardware-based:* Some systems have been proposed that use high-frequency wireless radio signals and special antenna alignment to improve the performance of human recognition systems. For example, in order to extract small Doppler shifts from OFDM Wi-FI transmissions to recognize human gestures, WiSee [10] uses USRP as wireless devices and utilizes communication on a 10 MHz channel at 5 GHz. Adib et al. proposed WiTrack [8], which leverages specially designed Frequency Modulated Carrier Wave (FMCW) to get accurate Time-of-Flight (ToF) measurements. In their settings, directional antennas, which are arranged in a "T", are also used in WiTrack to help recognize human gestures through walls.

*2) RSS-based:* Various systems use RSS collected from commercial Wi-Fi chipsets for human activity recognition [11] and human localization [12, 13]. Abdelnasser et al. proposed WiGest [11], which uses RSS waveforms to detect different gestures over the laptop. In SpotFi [12] (proposed by Kotaru et al.) and Wideo [13] (proposed by Joshi et al.), RSS is used to calculate the distance between the transmitter and the target. However, RSS values collected from commercial Wi-Fi devices only provide coarse-grained channel variation information. Furthermore, they cannot utilize multi-path effects of indoor Wi-Fi signals. As a result, most systems only use RSS for macro-movement recognition and distance estimation.

*3) CSI-based:* Compared with RSS, CSI can provide not only fine-grained channel status information, but information about small scale fading and multi-path effects caused by micro-movements. Most wireless signal-based systems use CSI values as data source, and their approaches either follow the structure of machine learning systems [1, 14, 15] or find some common patterns among different people [16, 17].

Ali et al. proposed Wikey [14], which can recognize keystrokes of different users in an indoor environment. The key intuition is that different keystrokes generate different CSI waveforms, and different waveforms can be used

as features. CARM [1], proposed by Wang et al., has two theoretical underpinnings: a CSI-speed model, which quantifies the correlation between CSI value dynamics and human movement speeds, and a CSI-activity model, which quantifies the correlation between the movement speeds of different human body parts and a specific human activity. By these two models, they quantitatively build the correlation between CSI value dynamics and a specific human activity. Wang et al. proposed WiHear [3], which recognizes mouth movements and "hears" people talk within the radio range. Han et al. proposed WiFall [18] which can recognize the fall of the target in an indoor environment. Li et al. proposed WiFinger [2], which can use ubiquitous wireless signals to achieve number text input in Wi-Fi devices by recognizing finger-grained gestures. The system designed in [19] can extract human gait information and individual specific features from spectrograms. These systems follow the general structure of machine-learning based systems and have four stages: noise removal, feature extraction, classification, and evaluation. Different feature extraction and classification models are used in these systems, such as KNN, SVM, HMM, and so on.

Some other systems are not based on machine learning. In [16], Zou et al. found that CSI values distribute more widely and change more drastically when there are more moving people. They designed Electronic Frog Eye to count the number of people in a crowd based on this observation. Sun et al. found that if the user's hand blocks a signal arriving along a specific Angle-of-Arrival (AoA), the RSS of this signal will experience a sharp drop. Then, we can localize hands by monitoring RSS changes of signals with different AoAs [17].

While most of the existing CSI-based human activity recognition systems focus on recognizing human activity to provide more human-computer interfaces, such as gestures, lip movements, and keystroke, Wi-Fi sensing technologies have the potential to be used in health monitoring and rescue situations. In addition to being influenced by macro-movements, Wi-Fi signals are influenced by micro-movements, such as chest movements. There are already some works which use CSI to monitor vital signs, such as [20] and [21].

## III. System Overview

### A. Challenges

To reduce the required number of labeled instances, three technical challenges need to be addressed in our system. The first challenge is how to use the knowledge in the labeled dataset to label those unlabeled instances. This is challenging since the labeled data is noisy and not enough to train a robust classifier. Besides, the unlabeled data may also contain some noisy samples, so not all the unlabeled data can be used as auxiliary data. Our solution is to use the labeled data to train multiple classifiers with multiple

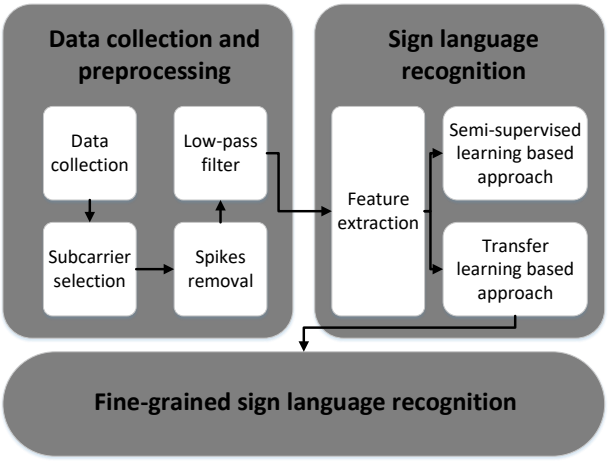



Figure 2. System structure

models. Then, we use these multiple classifiers to predict those unlabeled instances. An unlabeled instance is labeled as $y_i$ if and only if the predicted labels of all classifiers are $y_i$.

The second challenge is how to transfer the others' knowledge to the new user in order to train a new robust classifier under the new setting. In practice, human recognition systems could just get a few labeled instances of the new user without unlabeled instances, since the labeling progress tends to be inconvenient and expensive. Moreover, the others' labeled instances cannot be used as training instances for the new user directly. Due to the different floor plans and furniture placements, even the same human activity may introduce different multi-path distortions in Wi-Fi signals, and different human activities may still generate the same CSI waveform. In our system, we will calculate the similarity between new user's instances and others' instances based on their feature distributions and labels. We choose the labeled instances that are quite similar to existing labeled instances of the new user and add them into new user's training dataset. Our results show that the extended training dataset is large enough to train a robust classifier for the new user.

The third challenge is finding proper and efficient kernel functions. Due to a complex instance distribution on the feature hyperplane, it is usually hard to use a simple linear function to split the instances in SVM model. Based on the experimental results collected from our testbed, we define two kernel functions that can maximize the margin of the instances of different labels. Considering that our system supports three human gestures, we design a two-stage classification model with two kernel functions.

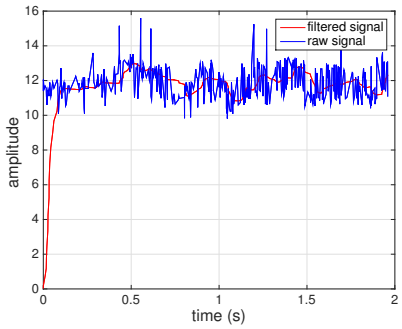



Figure 3. Noise removal

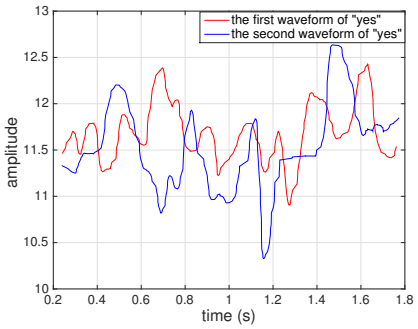



Figure 4. Two filtered CSI waveforms of "Yes"

### B. System structure

The main idea of our system is to recognize sign language based on CSI waveform analysis. The system flows are illustrated in Fig. 2. After we capture CSI waveforms from commercial Wi-Fi devices, we choose a subcarrier that is most sensitive to human gestures. We then apply a spikes removal filter on the raw signal to remove those samplings which are far away from their neighbor's samplings. Considering sign language is mostly low frequency, we apply a low-pass filter on the CSI waveform to remove noise which is at a high frequency. The filtered CSI waveform can then be used for feature extraction and classification.

Eight features are used in our system: the average amplitude, the maximal amplitude, the average median absolute deviation (MAD), the maximal MAD, the average normalized standard deviation (STD), the maximal STD, the average velocity, and the maximal velocity. Furthermore, not all of these eight features can be used to classify human gestures, so we further chose two features which can represent gestures effectively.

After feature extraction, each waveform $X_i$ (or instance) can be represented as a vector $X_i = (x_1, x_2, \cdots, x_n)$, where $x_i$ means a predefined feature. If the new user only has a few labeled instances, we will use the transfer learning-based method to improve the recognition performance with the help of auxiliary data. Here, auxiliary data is obtained from others' labeled instances.

## IV. Preprocessing

In this section, we describe how we choose proper subcarrier and remove noise from the raw signal.

### A. Subcarrier selection

The IWL5300 provides us 802.11n channel state information in a format that reports the channel matrices for 30 subcarrier groups. At each subcarrier, the fine-grained CSI describes how a signal propagates from the transmitter to the receiver with the combined effect of, for example, scattering, fading, and power decay with distance. Based on collected data, we find that different subcarriers have different sensitivities to different human activities. In order to obtain a robust sign language estimation, a proper subcarrier needs to be chosen. Considering 3 activities supported in our system, we choose CSI waveforms of the fifth subcarriers of the channel between the first transmit antenna and receive antenna.

### B. Noise removal

The raw CSI waveform collected from commercial Wi-Fi devices is usually noisy so that it cannot be used directly for feature extraction. In our system, we first use a median filter to smooth the waveform. Due to the poor performance of the median filter with high-frequency noise, we further apply a low-pass filter on the CSI waveform to remove the high-frequency component that cannot be caused by human hand and arm movements. Since 3 activities supported in our system are performed at a low frequency, these two filters can still effectively remove noise and keep useful information. Fig. 3 illustrates the effectiveness of our data preprocessing by comparing the CSI waveform before and after data preprocessing.

## V. Feature Extraction

Proper features are important for classification. In our experiments, we find that the patterns of two users may be not the same. Since the wavelength of the $2.4GHz$ Wi-Fi signal is about 12 centimeters, the shapes of two waveforms can be quite different. For instance, in Fig. 4, both waveforms represent "Yes", but it is not easy to find a common pattern between them. This means we cannot directly use the shape of the waveform as the feature.

To differentiate among different gestures, we try to find proper features that can uniquely represent different sign languages. In our system, we extract eight key waveform features from the filtered signal: (1) mean amplitude of filtered waveform, (2) maximal amplitude of filtered waveform, (3) the average median absolute deviation (MAD) value, (4) the maximal MAD value, (5) the average normalized standard deviation (STD) value, (6) the maximal STD value, (7) the

average velocity of the signal change, and (8) the maximal velocity of the signal change. However, not all 8 features can be used in the classification. We find that we cannot easily distinguish sign language based on some features, such as the average velocity of the signal change or the average STD. Based on our experimental results, the average amplitude of the filtered waveform and average MAD value are more useful for classification.

## VI. CLASSIFICATION

After extracting useful features from the raw CSI waveform, suitable classification models and methods should be determined for the best prediction. Here, we consider two cases. In the first case, the new user has a lot of data, but only a few are labeled. In the second case, the new user only has a few labeled instances, while our system has saved a lot of labeled instances from other users and different environments. Our system prefers using a semi-supervised learning-based solution for the first case, which is described in detail in Section VI-A. The transfer learning-based approach discussed in Section VI-B will address the second case.

### A. Semi-supervised learning-based solution

In many machine learning-based approaches adopted by existing HAR systems, the target function is estimated using labeled data. However, labeled instances are often very time consuming and expensive to obtain. The new user may only be able to label some instances, while most instances stay unlabeled. Semi-supervised learning aims to address this issue. Along with labeled instances, it exploits unlabeled ones to improve learning performance. Co-training is an efficient semi-supervised learning paradigm, which trains two classifiers through letting them label the unlabeled instances.

In our system, we adopt a similar idea which is used in classic co-training [22] and En-Co-training [23]. In order to reduce the cost on feature extraction, we only adopt one feature view here. Here we consider a basic binary classification example. In the beginning, all the labeled data will be used to train two classifiers based on SVM and KNN, respectively. These two classifiers are then used to predict possible labels of unlabeled instances. For instance $u_i$, if the predicted labels $y_i^1$ and $y_i^2$ of SVM and KNN are the same, then $u_i$ is labeled as $y_i^1$. Instance $u_i$ is then added to the labeled instances set to train the SVM and KNN classifiers again. The detailed algorithm is listed in Algorithm. 1.

The reason we use two classifiers here is to make the prediction more accurate. For instance, if we just use the SVM classifier, a small prediction error in the first iteration may cause irreparable errors after several iterations. By introducing a KNN classifier, we can avoid this to a large degree while still keeping a good prediction performance. Moreover, since different features have different units and ranges, we use normalized feature values to replace absolute feature values on all the dimensions.

**Algorithm 1** Semi-supervised learning-based: source code

**Input:** a set of labeled instances $L$, a set of unlabeled data $U$, label space $Y$, the number of iterations $k$

**Output:** a set of labeled instances, $L'$.

```
while k > 0 do
    use L to train two classifiers M_1 and M_2 based on SVM model and KNN model.
    for u_i ∈ U do
        use M_1 to predict u_i and get a predicted label y_i^1
        use M_2 to predict u_i and get a predicted label y_i^2
        if (y_i^1 = y_i^2) then
            label u_i as y_i^1 and move u_i to L
```

### B. Transfer learning-based solution

To use the transfer learning-based approach, we first need to find those useful instances from existing large labeled data (collected from other users). Here, useful instances mean those instances which can be directly used as labeled instances in the classification of a new user, and the set of useful instances is called the auxiliary set.

Given two instances $x_i$ and $x_j$ with labels $y_i$ and $y_j$, we need to determine whether these two instances are quite similar to each other. Since all the feature values we have are continuous, it is hard to find two instances that are the same as each other. Moreover, since the average amplitude and average MAD have different units and ranges, we cannot use the geometric distance directly to measure the distance of two instances on the hyperplane. Based on these observations, we discretized all the feature values in both dimensions. The feature value after discretization on each dimension is calculated by the following equation:

$$F_d(i,j) = \lceil (F(i,j) - Min(i))/\tau \rceil$$
$$i = 1, \ldots, N_f \quad j = 1, \ldots, N_s$$

where $N_f$ is the number of selected features, $N_s$ is the number of instances, $Min(i)$ represents the minimum value of all instances on the $i^{th}$ features, $F(i,j)$ is the absolute value of the $j^{th}$ sample for the $i^{th}$ feature, and $F_d(i,j)$ is the absolute value of $j^{th}$ sample for the $i^{th}$ feature after discretization. Here, $\tau$ is the unit of discretization.jjjjj If the discrete feature values are the same on all dimensions for two instances and their labels are the same, then we argue that these two instances are quite similar. We use this method to find all similar instances in the others' training dataset and include those instances into the training dataset of the new user. The knowledge transfer results are illustrated in Fig. 5(a). We can observe that our method can increase the number of training instances effectively and does not change the original instance's distribution.

Here, we use the same formulation of the SVM classifier with the auxiliary dataset in [24]. A typical SVM has the following form:

$$y = \begin{cases} 1, & \sum_j \alpha_j y_j K(x_j, x) + b \geq 0 \\ 0, & otherwise \end{cases}$$

where the $\alpha_j$ and $b$ are the learned parameters and the function $K(x_j, x)$ is a kernel function that we designed based on the instance distribution on the feature hyperplane.

The values of both $\alpha$ and b are learned by solving a convex optimization problem. In this paper, we consider a linear programming SVM since it encourages sparser solutions than the usual SVM quadratic regularization penalty.

$$\begin{aligned} \text{Minimize} \quad & \sum_j \alpha_j + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i(\sum_j y_j \alpha_j K(x_j, x_i) + b) + \xi_i \geq 1 \quad \forall i \\ & \alpha_j \geq 0 \quad \forall j \end{aligned}$$

The $\sum_j \alpha_j$ penalizes the complexity of the classifier, and the $C \sum_i \xi_i$ measures how poorly the classifier fits the training data. The slack variables $\xi_i$ will be positive precisely for those training examples where the classifier does not classify correctly with a margin of at least 1.

For the instances in the auxiliary dataset, they can be used as support vectors or included in constraints. Those instances with index $i$ are used as support vectors, and instances with index $j$ represent those that are included in constraints. Then, we will have following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_j^{N^p} \alpha_j^p + \sum_j^{N^a} \alpha_j^a + C^p \sum_i^{N^p} \xi_i^p + C^a \sum_i^{N^a} \xi_i^a \\ \text{subject to} \quad & y_i^p(\sum_j^{N^p} y_j^p \alpha_j^p K(x_j^p, x_i^p) + sum_j^{N^a} y_j^a \alpha_j^a K(x_j^a, x_i^p) \\ & + b) + \xi_i^p \geq 1 \quad i = 1, \ldots, N^p \\ & y_i^p(\sum_j^{N^p} y_j^p \alpha_j^p K(x_j^p, x_i^a) + sum_j^{N^a} y_j^a \alpha_j^a K(x_j^a, x_i^a) \\ & + b) + \xi_i^a \geq 1 \quad i = 1, \ldots, N^p \\ & \alpha_j^p \geq 0 \quad j = 1, \ldots, N^p \\ & \alpha_j^a \geq 0 \quad j = 1, \ldots, N^p \end{aligned}$$

As illustrated in Fig. 5(a), although we have included auxiliary data, it is still hard to use classic linear SVM to classify all the instances in the feature hyperplane. Moreover, traditional SVM classifiers only support two labels, so more than one SVM classifier should be trained to distinguish three sign languages. In our system, we use a two-stage classification with three classifiers to recognize these three sign languages. Firstly, we use a Polynomial function as the kernel function to classify "Thanks", and the result is illustrated in Fig. 5(b). Then, we remove all the instances that are labeled as "Thanks" from the training set. For the other two sign languages, we also adopt a polynomial function as the kernel function to classify them. The classification contour and support vectors are illustrated in Fig. 5(c).

## VII. Evaluation

In this section, we describe our hardware setup and data collection of WiSign. Then, system performance is well evaluated different settings.

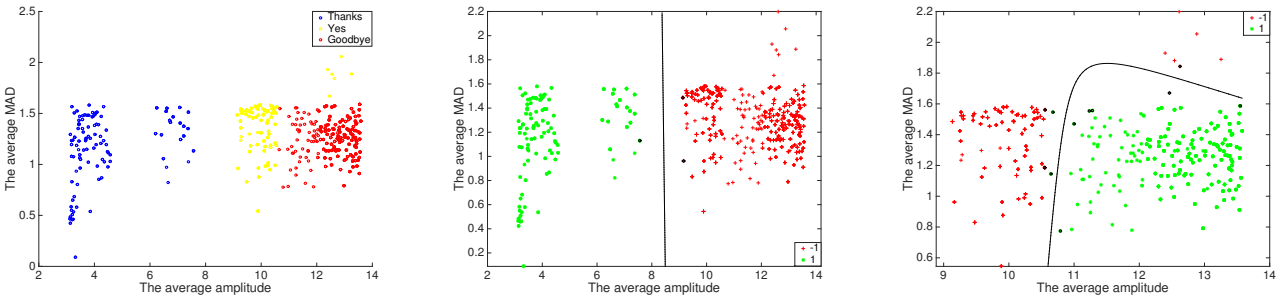
### A. Hardware setup

We implement our system using two Commercial off-the-shelf (COTS) Wi-Fi devices. Specifically, we use a Lenovo X210 laptop with Intel Link 5300 Wi-Fi NIC as the receiver to record the CSI measurements. The laptop has a 2.13 GHz Intel Core™ I3 processor with 2GB of RAM and Ubuntu 14.04 as its operating system. We use a TP-Link TL-WR1043ND Wi-Fi router as the transmitter and set the router in AP mode at 2.4 GHz. Since the modified driver can only get CSI measurements from $802.11n$ packets, we modify the network configuration to make sure that packets are sent under the $802.11n$ protocol. To increase the sampling rate, we set up an FTP server on a Macbook pro laptop in the same local area network and let the receiver continuously download a large file via the transmitter. Based on the experimental results, the average sampling rate of our system is about 125 CSI measurements per second. All the CSI measurements are collected from Intel 5300 NIC using a modified driver developed by Halperin et al. [25]. In our testbed, we have three antennas for the transmitter and two linearly assigned antennas for the receiver, so we can get totally $3 \times 2 \times 30 = 180$ different CSI waveforms on different subcarriers for each measurement.

### B. Data collection

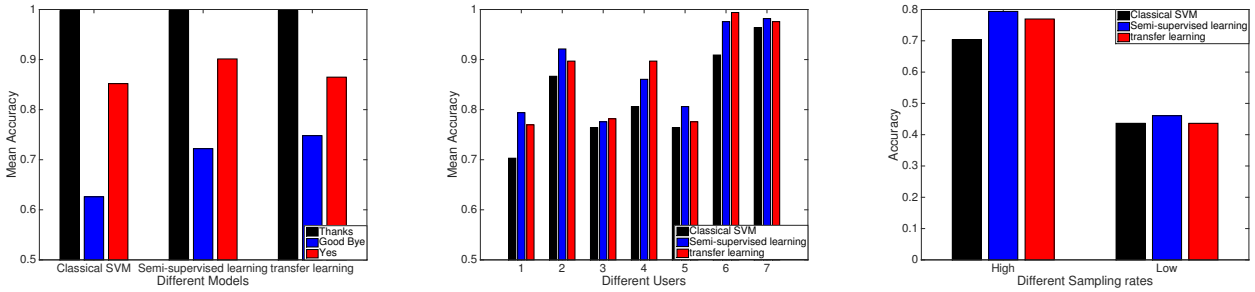
To evaluate the performance of the transfer learning-based and the semi-supervised learning-based approaches proposed in our system, we collected training data and testing data from seven users in the same room. None of these users have experience or knowledge on Wi-Fi signal-based human recognition systems before these experiments. The Wi-Fi transmitter and receiver are placed in a straight line on a desk at a distance of about 0.2 meters. The distance between the participant and the receiver is about 0.2 meters during all experiments. The FTP server is placed on the other side of the receiver on the same straight line at a distance of about 0.13 meters. Each participant is asked to repeat each gesture at least 60 times under different experiment settings. The locations of all the furniture and devices are not changed during all experiments.

We evaluate the sign language prediction performance of WiSign for our two solutions and compare their results with the traditional training method that only uses the SVM learning model with the kernel function. To evaluate the




(a) Instances distribution after knowledge transfer (b) Classification result for "Thanks" (c) Classification result for "Good bye" and "Yes"

Figure 5. Transfer learning-based solution




(a) Influences of different models (b) Influences of different participants (c) Influences of different sampling rates

Figure 6. Evaluation results

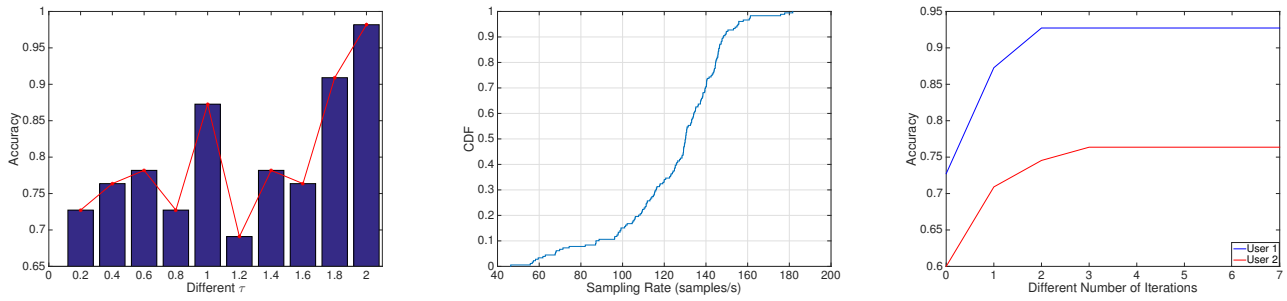
performance of the transfer learning-based approach, 15 labeled instances (5 instances for "Good Bye", 5 instances for "Thanks", and 5 instances for "Yes") from the dataset of one participant are used as the primary training dataset. The other labeled instances from the other six participants are used as the possible auxiliary dataset. To evaluate the performance of the semi-supervised learning-based solution, we only use instances from one user. The same 15 instances are used as the labeled dataset, $L$, in Algorithm 1, and the rest of the instances are used as the unlabeled dataset. For comparison, we use the same 15 instances to train a classic SVM classifier without leveraging extra knowledge. We further evaluate the influence of different $\tau$ and the number of iterations in our two approaches. We also study the distribution of the sampling rate in our testbed and explore the system's performance under an extremely low sampling rate.

### C. Prediction accuracy vs. different classification methods

In this experiment, we use the same initial training dataset for the three approaches and explore their performances on the same testing dataset. Fig. 6(a) illustrates the prediction performance of our two solutions and the classic SVM approach on one user. We can observe that both the transfer learning-based solution and the semi-supervised learning solution achieve better prediction accuracies than the classic SVM approach. Since the instance cluster of "Thanks" is far away from the other two clusters on the feature hyperplane, all three methods can achieve a good prediction performance of $99.74\%$. For " Good Bye" recognition, our two methods can achieve a mean prediction accuracy of $72.21\%$ and $74.81\%$, respectively. While the classic SVM classifier can only provide a low prediction performance of $62.5\%$, our two methods improve the accuracies of "Yes" prediction by $4.94\%$ and $1.3\%$, respectively. These results show that our semi-supervised learning-based and transfer learning-based methods can improve the prediction accuracy by leveraging the knowledge from unlabeled data and the others' training dataset.

### D. Mean prediction accuracy vs. different participants

We further studied the mean prediction accuracy of WiSign among all involved users. We can still observe from Fig.6(b) that our two solutions still have a better mean prediction accuracy compared with the classic SVM method. More specifically, our transfer learning-based approach and semi-supervised learning-based approach achieve the mean prediction accuracies of $79.39\%$ and $76.97\%$ for user 1, while the mean accuracy of the classic SVM is only $70.3\%$. Since the patterns of the first user's activities are quite different from the other 6 users, only limited knowledge can be leveraged by our methods. Similarly, the instance




(a) Influences of different $\tau$ (b) Distribution of sampling rates (c) Influences of different iterations

Figure 7. Evaluation results

distributions of the third and the fifth users differ from the other users, but our methods still improve the prediction accuracy by at least $1.22\%$. The mean accuracies of our solutions for the second user are $92.12\%$ (transfer learning-based approach) and $89.7\%$ (semi-supervised learning-based approach), while the classic SVM has the mean accuracy of only $86.67\%$. In some cases, the semi-supervised learning-based approach may have better results than the transfer learning-based approach due to the instance distribution on the feature hyperplane. For example, for user 7, the semi-supervised learning-based approach has a mean accuracy of $98.18\%$, which is better than that of transfer learning ($97.58\%$). If two users have the very similar pattern of supported activities, the transfer learning-based approach can achieve a better performance. For the sixth user, the transfer learning based approach can achieve a mean accuracy of $99.39\%$, which is better than that of the semi-supervised learning based approach ($97.58\%$). The results show that WiSign can really improve the recognition accuracy even if the participants have different speeds and ranges of same gesture.

### *E. Mean prediction accuracy vs. different sampling rates*

In this subsection, we evaluated the influence of different CSI sampling rates. Instead of downloading a large file from an FTP server, we let the receiver keep pinging the transmitter every 0.05 seconds. Based on the experimental results, the average sampling rate is about 23.2 samples per second. Under a low sampling rate, most waveform details have been lost, and the extracted features can be easily influenced by noisy measurements. We collected the dataset under a low sampling rate from the first user with the same size of the dataset and compared the mean accuracy with that under a high sampling rate. We observe from Fig. 6(c) that the mean prediction accuracy drops rapidly when the sampling rate decreases. The mean prediction accuracies of the transfer learning-based approach and semi-supervised approach decrease to $46.06\%$ and $43.64\%$, respectively. The high sampling rate can provide us with more information about the channel state within the same time duration. In the future, we will try to improve the sampling rate to more than 2000 samples per second and study whether the mean accuracy will further increase.

### *F. Prediction accuracy vs. different $\tau$*

In transfer learning-based approach, how to determine the value of $\tau$ is a serious issue. In our experiments, we evaluate the recognition performance of "Good Bye" of the sixth user under different $\tau$, and the results are illustrated in Fig. 7(a). We can observe that the recognition performance does not always improve with the increase of $\tau$. For example, the waveform fluctuates between $69.09\%$ and $87.27\%$ when $\tau < 1.6$, but most accuracies are still better than that without knowledge transferred ($\tau = 0$). The prediction accuracy keeps increasing when $1.6 \leq \tau \leq 2$. This is because more useful instances in others' training data set can be properly used for local classification. When $\tau = 2$, the prediction accuracy reaches $98.18\%$, which means our system is robust enough in practice.

### *G. Sampling rate distribution*

To evaluate the sampling rate performance of our data collection, we study the sampling rate distribution across all the collected data. We present the distribution using Cumulative Distribution Function (CDF) graph that is illustrated in Fig. 7(b). We can see that more than $90\%$ instances have sampling rates that are higher than 80 samples per second. At least half of the total data has sampling rates that are higher than 130 samples per second. Most of the sampling rates are between 90 and 160 samples per second. The results show that ost of our collected data has enough sampling rates to provide sufficient information for following classification and prediction.

### *H. Prediction accuracy vs. different number of iterations*

There is a trade-off on determining the number of iterations. If the number of iterations is small, we can reduce the overhead of data processing, while limited unlabeled data can be used for a better classification. If the number of iterations is too large, we can ensure a good performance

from our semi-supervised learning-based method, but we need more computing resources. In order to estimate what may be the best number of iterations, we study the influence of a different number of iterations on prediction accuracies of 2 users, which is shown in Fig. 7(c). We can observe that the recognition accuracy increases with the increase of the number of iterations for both users at the beginning. After that, the recognition performance keeps steady no matter how many iterations we run for. Based on these experimental results, we set the number of iterations as 5 to get the best recognition performance while reducing the computing overhead.

## VIII. CONCLUSION

In this paper, we propose WiSign, a Wi-Fi signal-based indoor sign language recognition system. We propose two approaches based on transfer learning and semi-supervised learning to reduce the required number of labeled instances in the classification stage. In the transfer learning-based solution, existing similar knowledge from others' labeled datasets are used to act as the auxiliary dataset. In the semi-supervised learning-based solution, we exploit unlabeled instances to improve the learning performance. We implement our system using a Lenovo X210 laptop with Intel Link 5300 Wi-Fi NIC as the receiver and a TR-Link TL-WR1043ND as the transmitter. Our experimental results show that WiSign can achieve the mean prediction accuracies of $87.01\%$ (the transfer learning-based approach) and $87.38\%$ (the semi-supervised learning-based approach) for all participants.

## ACKNOWLEDGMENT

This research was supported in part by NSF grants CNS 1629746, CNS 1564128, CNS 1449860, CNS 1461932, CNS 1460971, CNS 1439672, CNS 1301774, ECCS 1231461, and ECCS 1231461.

## REFERENCES

[1] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the MobiCom*. ACM, 2015, pp. 65–76.

[2] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "Wifinger: talk to your smart devices with finger-grained gesture," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 250–261.

[3] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014, pp. 593–604.

[4] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Wiwho: wifi-based person identification in smart spaces," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE Press, 2016, p. 4.

[5] T. Huynh and B. Schiele, "Towards less supervision in activity recognition from wearable sensors," in *Proceedings of the ISWC*. IEEE, 2006, pp. 3–10.

[6] T. S. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, pp. 487–493, 1999.

[7] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[8] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3d tracking via body radio reflections," in *Proceedings of the NSDI*, 2014, pp. 317–329.

[9] *http://www.castlerockfamilyenrichmentcenter.com/workshops/baby-signing/*.

[10] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the MobiCom*. ACM, 2013, pp. 27–38.

[11] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *Proceedings of the INFOCOM*. IEEE, 2015, pp. 1472–1480.

[12] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proceedings of the SIGCOMM*. ACM, 2015, pp. 269–282.

[13] K. Joshi, D. Bharadia, M. Kotaru, and S. Katti, "Wideo: Fine-grained device-free motion tracing using rf backscatter," in *Proceedings of the NSDI*, 2015, pp. 189–204.

[14] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using wifi signals," in *Proceedings of the MobiCom*. ACM, 2015, pp. 90–102.

[15] J. Shang and J. Wu, "A robust sign language recognition system with multiple wi-fi devices," 2017.

[16] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang, "Electronic frog eye: Counting crowd using wifi," in *Proceedings of the INFOCOM*. IEEE, 2014, pp. 361–369.

[17] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "Widraw: Enabling hands-free drawing in the air on commodity wifi devices," in *Proceedings of the MobiCom*. ACM, 2015, pp. 77–89.

[18] C. Han, K. Wu, Y. Wang, and L. M. Ni, "Wifall: Device-free fall detection by wireless networks," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 271–279.

[19] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 363–373.

[20] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *Proceedings of the MobiHoc*. ACM, 2015, pp. 267–276.

[21] J. Shang and J. Wu, "Fine-grained vital signs estimation using commercial wi-fi devices," in *Proceedings of the Eighth Wireless of the Students, by the Students, and for the Students Workshop*. ACM, 2016, pp. 30–32.

[22] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the COLT*. ACM, 1998, pp. 92–100.

[23] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee, "Activity recognition based on semi-supervised learning," in *Proceedings of the RTCSA 2007)*. IEEE, 2007, pp. 469–475.

[24] P. Wu and T. G. Dietterich, "Improving svm accuracy by training on auxiliary data sources," in *Proceedings of the ICML*. ACM, 2004, p. 110.

[25] *http://dhalperi.github.io/linux-80211n-csitool/*.