A Robust Sign Language Recognition System with Multiple Wi-Fi Devices

Jiacheng Shang and Jie Wu
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122
jiacheng.shang@temple.edu,jiewu@temple.edu

ABSTRACT

Sign language is important since it provides a way for us to the deaf culture and more opportunities to communicate with those who are deaf or hard of hearing. Since sign language chiefly uses body languages to convey meaning, Human Activity Recognition (HAR) techniques can be used to recognize them for some sign language translation applications. In this paper, we show for the first time that Wi-Fi signals can be used to recognize sign language. The key intuition is that different hand and arm motions introduce different multi-path distortions in Wi-Fi signals and generate different unique patterns in the time-series of Channel State Information (CSI). More specifically, we propose a Wi-Fi signal-based sign language recognition system called WiSign. Different from existing Wi-Fi signal-based human activity recognition systems, WiSign uses 3 Wi-Fi devices to improve the recognition performance. We implemented the WiSign using a TP-Link TL-WR1043ND Wi-Fi router and two Lenovo X100e laptops. The evaluation results show that our system can achieve a mean prediction accuracy of 93.8% and mean false positive of 1.55%.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; *Mobile devices*; • **Computing methodologies** → **Supervised learning**; **Support vector machines**;

KEYWORDS

Human activity recognition, Machine learning, Signal processing, Wi-Fi signals

1 INTRODUCTION

Sign language is important since it provides a way for us to the deaf culture and more opportunities to communicate with those who are deaf or hard of hearing. Some gestures defined in American Sign Language (ASL) are shown in Fig. 1. Since sign language chiefly uses body languages to convey meaning, Human Activity Recognition (HAR) techniques can be used to recognize them for some sign language translation applications. Traditional approaches for recognizing human activities can be categorized into three groups: camera-based approaches, low-cost radar-based approaches, and wearable sensor-based approaches. However, all of these traditional approaches have some limitations. Camera-based approaches need a line-of-sight and sufficient lighting and sometimes they may breach human privacy (e.g. in the bathroom). Low-cost radar-based systems have limited operation ranges of just tens of centimeters. Wearable sensor-based solutions require users to wear their sensors all the time, which is inconvenient and not practical in some applications (e.g. rescue applications).

In the past few years, several Wi-Fi signal-based approaches have been proposed to recognize human activity. The key intuition is that different human activities will introduce different multi-path distortions in Wi-Fi signals and generate a unique pattern in the time-series of Channel State Information (CSI) values. Some commercial Wi-Fi devices provide us a fine-grained CSI in time-series. Because of the high data rate provided by these modern commercial Wi-Fi devices, we can get enough samples of CSI measurements within the duration of human activities. Various human activities have also been supported in these systems. For instance, Ali et al. proposed Wikey [5] which can recognize keystrokes of different users in an indoor environment. Wang et al. proposed WiHear [9] that can recognize mouth movement and "hear" people talk within a radio range. Han et al. proposed WiFall [4] which can recognize the falling event of a user in the indoor environment. These systems follow the general structure of machine learning-based systems and generally have four stages: data collection, noise removal, feature extraction, and classification. Different feature extraction and classification models are used in these systems, such as Principal Component Analysis (PCA), anomaly detection, Discrete Wavelet Transform (DWT), and Hidden Markov Model (HMM).

Most of activities supported in existing systems can be distinguished based on the CSI measured from one receiver. Howerver, some human activities like gestures defined in ASL have the similar frequency and will introduce almost the same multi-path distortions in Wi-Fi signals for one Wi-Fi receiver. In this case, existing Wi-Fi signal-based HAR systems tend to have bad performance. Moreover, most of existing Wi-Fi-based approaches did not discuss the false positive of recognition. For our sign language recognition system, we want to reduce the false positive of recognition since we want to include as less noisy activities as possible.

Figure 1: Sign language [2]

In this paper, we propose a Wi-Fi signal-based sign language recognition system, called WiSign. Different from existing Wi-Fi signal-based HAR systems, WiSign consists of one Wi-Fi router and two receivers. The router keeps emitting signals, and two receivers receives those signals and record CSI measurements. When a user performs a specific hand or arm movement within the range of WiSign, each receiver recognizes the activity based on the analysis of the variety of CSI waveforms. After getting the recognition results at each side, two receivers exchange their prediction results and do a weighted majority voting based on the prediction confidence. Based on the experiment results, the solutions we proposed can reduce the false positive and increase recognition accuracy, which makes our Wi-Fi-based sign language recognition system more practical in the real scenarios.

The remainder of this paper is organized as follows: In Section 2, we will introduce current HAR system using special hardware, Received Signal Strength (RSS), and CSI. In Section 3, we will discuss the challenges we faced and the structure of WiSign. Signal preprocessing, feature extraction, and classification algorithms will be discussed in Sections 4, 5, and 6. In Section 7, we will introduce our experiment implementation and analyze the evaluation results. The final conclusion and future work will be given in Section 8.

2 RELATED WORK

2.1 Special hardware-based

Some researchers proposed some systems that use high-frequency wireless radio signals and special antenna alignment to improve the performance of human recognition systems. WiSee [7] uses USRP as wireless devices and utilizes communication on a 10 MHz channel at 5 GHz. This implementation can help extract small Doppler shifts from OFDM Wi-FI transmissions to recognize human gestures. WiTrack [4], proposed by Adib *et al.*, uses specially designed Frequency Modulated Carrier Wave (FMCW) to get accurate Time-of-Flight (ToF) measurements. Directional antennas, which are arranged in a "T", are also used in WiTrack to help recognize human gestures through walls.

2.2 RSS-based

The RSS collected from commercial Wi-Fi chipsets can be used for human activity recognition [3] and human localization [6]. Abdelnasser *et al.* proposed WiGest [3] which uses RSS waveform to detect different gestures over the laptop. In SpotFi [6] proposed by Kotaru *et al.*, they use RSS to calculate the distance between the

transmitter and the target. However, RSS collected from commercial Wi-Fi devices only provide coarse-grained channel variation information. Furthermore, these systems can not utilize multi-path effects of indoor Wi-Fi signals. As a result, most systems only use RSS for macro-movement recognition and distance estimation.

2.3 CSI-based

Compared with RSS, CSI provides not only fine-grained channel status information, but information about small scale fading and multi-path effects caused by micro-movement. Most human recognition systems use CSI values as data source, and their approaches can also be divided into 2 categories: machine-learning based [5, 10] and non-machine-learning based [6, 11].

Ali *et al.* proposed Wikey [5] which can recognize keystrokes of different users in the indoor environment. The key intuition is that different keystrokes generate different CSI waveform, and different waveforms can be used as features. Wang *et al.* proposed WiHear [9] that can recognize mouth movement and "hear" people talks within the radio range. Both of these two systems are designed to recognize micro-movement, so their solutions and system settings cannot deal with hand and arm movement recognition well. Han *et al.* proposed WiFall [4] which can recognize the fall of the target in the indoor environment. However, only one activity (falling down) is supported in their system. CARM [10], 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. CARM recognizes different human activities based on the different frequencies. However, gestures defined in ASL share similar frequency and operation range, which makes it hard to distinguish them using frequency profile. 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 PCA, DWT, and HMM. Some other systems are not based on machine learning. In [11], 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. In SpotFi [6], they incorporate super-resolution algorithms that can accurately compute the angle of arrival (AoA) and use AoA to localize objects. These system are desinged mostly for counting people or localization, which makes them unsuitble for sign language recognition.

3 SYSTEM DESIGN

3.1 Challenges

To reduce the false positive of our WIi-Fi-based sign language system, two main challenges need to be addressed in our solutions. The first challenge is how to train two independent classifiers on two Wi-Fi devices. This is challenging since the two classifiers can make same mistake if they use a simialr training dataset, then the performance of our system cannot be really improved even if we
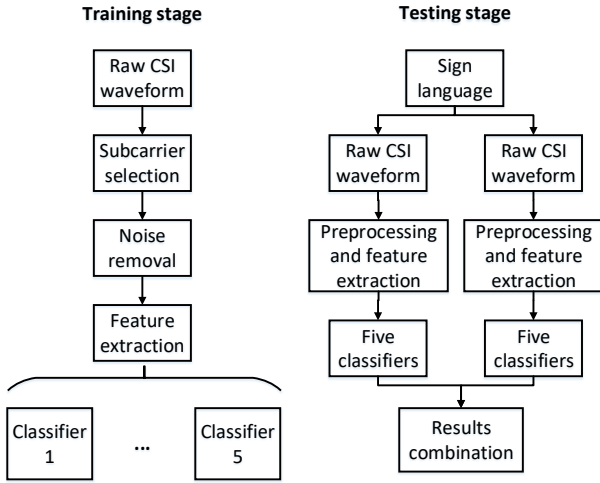
Training stage

Raw CSI waveform
↓
Subcarrier selection
↓
Noise removal
↓
Feature extraction

Classifier 1   ···   Classifier 5

Testing stage

Sign language
↓
Raw CSI waveform   Raw CSI waveform
↓   ↓
Preprocessing and feature extraction   Preprocessing and feature extraction
↓   ↓
Five classifiers   Five classifiers
↓   ↓
Results combination

Figure 2: System structure

combine the prediction results from multiple Wi-Fi devices. In our experiments, we explore possible locations of our system relative to the user and find the best location where the two Wi-Fi devices can receive totally different CSI profiles for the same human activity.

The second challenge is how to combile the prediction results properly on two Wi-Fi devices. The naive way is to consider each receiver as a player in majority voting, but we cannot guarantee we can always get a voting results even if we have odd number of receivers. In this paper, we train multiple classifiers on each receiver, and each of them will vote as a player based on their prediction confidences.

3.2   System Structure

The system flows are illustrated in Fig. 2. In the training stage, each Wi-Fi device collects the raw CSI waveform from the modified driver and removes the high-frequency noise. Since we care about the whole trend of CSI waverform, we further perform a smoothing filter on the processed CSI waverform. After preprocessing, we extract useful features from the filtered CSI waveform. Since CSI waveforms of different gestures differ most on the average amplitude and the average MAD, then, each instance (CSI waveform) is represented as a point in the two-dimenssion feature hyperplane, and we use SVM classifiers with different kernels to build classifiers on each of the Wi-Fi devices.

In the prediction stage, the CSI waveform in the moving window will be first processed with the same filters and feature extraction methods. Then, two waveforms which are casued by the same human gestures will be classified by each classifier, respectively. Based on the confidences (or scores) of these predictions, we perform a weighted voting and get the final prediction result.
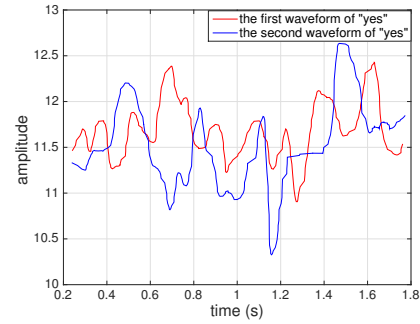

Figure 3: Two waveforms of "Yes"

## 4   PREPROCESSING

4.1   Subcarrier selection

In order to obtain a robust sign language estimation, we need to choose the subcarrier which is the most sensitive to sign language. The IWL5300 provides 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. We notive clearly that different waveforms have different sensitivities to sign language. Based on our experiment results, we always choose the CSI waveform that has the largest average amplitude between the first transmitting antenna and the first receiving antenna.

4.2   Noise removal

The raw CSI waveform we collect from commercial Wi-Fi devices is usually not reliable enough to be used for feature extraction because of the noise. The noise can be from environmental changes, radio signal interference, etc. In our system, we first use a median filter to mitigate the samples which have a significantly different value from other neighboring samples in a raw CSI waveform. Next, we apply a low-pass filter on the CSI waveform to remove high-frequency noise that cannot be caused by human hand and arm movement, due to the poor performance of the median filter with high-frequency noise. Since human sign language is at low frequency, these two filters can still effectively remove noise and keep useful information.

5   FEATURE EXTRACTION

In the real sign language recognition environment, it is hard to guarantee that the user always perfoms the same activity every time. Considering the speeds of human gestures may be not always the same even for same person, we will not always get exactly the same CSI waveform shape for the same human gesture. However, we can observe from Fig. 3 that the CSI waveforms of the same gesture will follow the same trend. For example, the average amplitude of these two waveforms are about 11.5. Since they also have the same trend, they should have similar average median absolute deviation value. Based on this observation, we test 8 different features of CSI waveforms: 1) mean amplitude of filtered waveform, 2) maximal
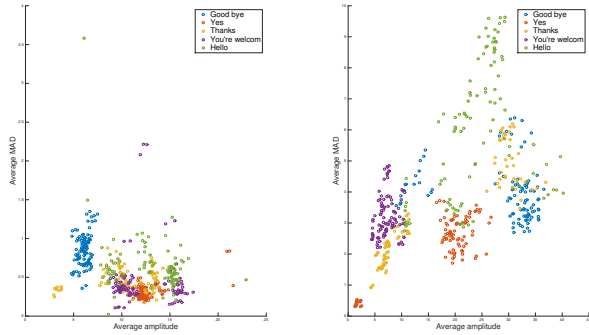
(a) Instances distributions on the first re-
ceiver

(b) Instances distributions on the second
receiver

Figure 4: Instances distributions on different receivers

Figure 5: Testbed

amplitude of the 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, 8) the maximal velocity of the signal change. However, not all of these 8 features are useful enough to be used in the classification. We find that it is hard to distinguish different sign language based on some features, such as the average velocity of the signal change and STD value. Based on our experiment results, the average amplitude and average MAD value of the filtered waveform are more useful for classification

Fig. 4 illustrates the instance distribuitons on two receivers in our testbed. We can observe clearly that "Good bye" can be perfectly classified. However, we cannot guarantee recognition performance of the other 4 sign language since their distributions have overlaps. We can see the same problem on the second receiver. On the second receiver, we can ensure good recoginition accuracy for "Yes", but points around $(34, 4)$ are hard to predict. They can be "Good bye", "Hello", or "Thanks".

6 CLASSIFICATION

6.1 Classifier training on one Wi-Fi Device

In our experiments, we collected the CSI waveforms of five different gestures ("Goodbye", "Hello", "You're welcom", "Thanks", and "Yes") that are defined in ASL. We find that for some activities, like "Good bye" on the first receiver and "Yes" one the second receiver, we can use the kernel-based SVM model to perfectly classify them. However, for some other activities, there are some overlaps among their distributions on given feature hyperplane. Without introducing more useful features and more powerful classfication models, the recognition accuracy of these gestures tends to be bad if we only use the data collected from one receiver. In our system, we have two laptops, and we will show that some gestures that are hard to classify on one laptop are easier to classify on another laptop. In our system, we adopt the SVM model with different kernel functions to classify these gestures. In order to solve our multiclass problem, we use an One vs. All (OVA) approach to classify each gesture on each receiver. Based on the instance distribution, we use

either the Guassian function or polynomial function as the kernel function to classify the objective gesture as perfectly as possible. These classifers are still not perfect, but they can provide useful information for our prediction combination stage.

6.2 Prediction Results Combination

We find that some activities whose distributions have overlaps on the feature hyperplane still cannot be perfectly classified on a single laptop. In our system, we improve the recognition performance for these "difficult gestrues" by adopting a weighted voting system. Weighted voting systems are voting systems based on the idea that not all voters should have the same amount of influence over the outcome of an election. Instead, it can be desirable to recognize differences by giving voters different amounts of say (mathematical weights) concerning the outcome. This is in contrast to a normal parliamentary procedure, which assumes that each member's vote carries equal weight. A weighted voting system is characterized by three things: the players, the weights and the quota. In our system, the players are the trained classifiers on both of these two laptops. The reason why we do not use each laptop as the player is that sometimes more than one classifiers on one laptop will have a prediction with a high confidence. If we always choose the result with the highest confidence, it may be a problem for those points that are in the middle of the overlaped region. The weights are the scores produced by each classifier itself. The quota is the minimum number of votes required to pass a motion. The returned scores of our SVM classifiers are from -1 to 1. The positive scores show how likely observations are classified in the positive class. The negative scores show how likely observations are classified in the negative class. Considering we do not know how negative scores will influence our combined prediction results, we defined a threshhould τ. In our system, we set the weights of those whose original weights are less than a threshhold τ to $W(\tau)$. If $\tau \geq 0$, $W(\tau) = 0$. Otherwise, $W(\tau) = \tau$. Then, we combine all the voting results together and choose the label with the largest combined weights.

Now we still have two problems for the results combination part. The first problem is that we cannot guarantee that we will always get the voting result. It is possible that the test instance is away from all the decision boundary we have trained and the biggest number of voting result is still less than the quota. In our system
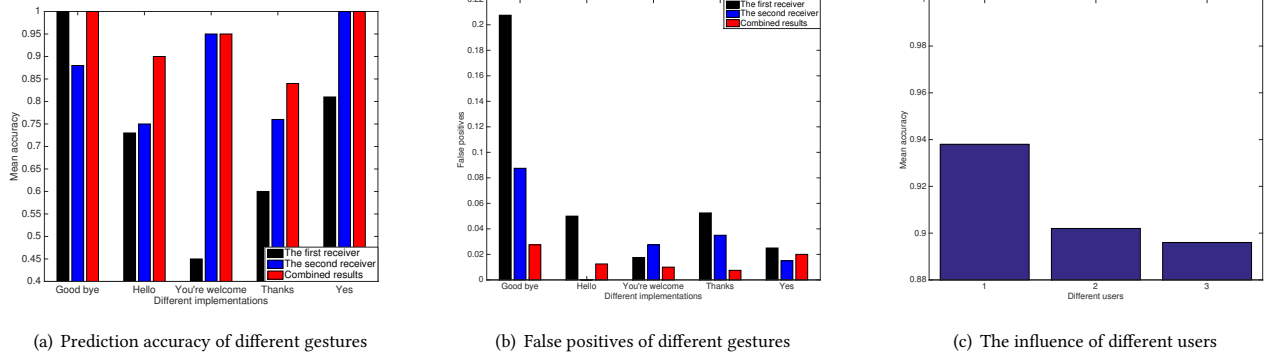
(a) Prediction accuracy of different gestures   (b) False positives of different gestures   (c) The influence of different users

Figure 6: Evaluation results

we will recognize this kind of instances as random noise that is not recoded in our training dataset. The second problem we have here is knowing when we should combine the prediction results. This is difficult to solve since we have no idea when the human gestures will occur and there is no synchronization between these two laptops. To address this problem, we assign different roles for the two laptops in our system. one of them acts as the primary node, and the other one acts as the secondary node. Initially, both the primary node and secondary node will predict a real-time CSI waveform frame every 0.1 second. Once at least one classifier of the secondary node has a prediction score larger than a threshhold τ, the secondary node sends all the prediction results and scores of the CSI waveform frame to the primary node. When the primary node gets the results and scores from the secondary node, it combines all the results and gives the final prediction. The next CSI waveform frame that is going to be analyzed will start from the last sample point of the previous frame that has the final voting result. The two receivers in our system can work as both primary node and secondary node. For example, Assume the prediction vector reported by the two vectors are (0.1, 0.2, 0.13, 0.78, 0.9, 0.27) and (0.12, 0.34, 0.2, 0.87, 0.14, 0.24), and the threshhould is set to 0.45. After removing those useless players, we get (0, 0, 0, 0.78, 0.9, 0) and (0, 0, 0, 0.87, 0, 0). We can see that the first laptop cannot distinguish the fourth and the fifth gesture. If we always choose the result with the highest confidence on one laptop, it is going to be a wrong prediction. But if we combine these two prediction vectors on the first laptop, we can get (0, 0, 0, 1.77, 0.78, 0). Based on the final combined prediction vector, we can get the correct prediction (fourth gesture).

7 EVALUATION

7.1 Hardware setup

We implement our system using 3 commercial Wi-Fi devices. Specifically, we use 2 Lenovo X210 laptops with Intel Link 5300 Wi-Fi NIC as the receivers. Each laptop has a 2.13 GHz Intel Core I3 processor with 2GB of RAM and Ubuntu 12.04 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. All the packets are sent under the 802.11n protocol. To increase the sampling rate, we set up an

FTP server on another laptop in the local area network and let 2 receivers continuously download a large file via the transmitter. Based on the experiment results, the average sampling rate of our system is about 267 samples per second. All the CSI samples are collected from Intel 5300 NIC using a modified driver developed by Halperin *et al*. [1]. We have three antennas for the transmitter and two linearly assigned antennas for the receiver, so we can get $3 \times 2 \times 30 = 180$ different CSI waveforms for each sample.

All experiments are conducted in an office room. Three different participants are involved over a two-month time period. As shown in Fig. 5, we implement our system on the table in a small meeting room. The distance between two laptops is about 0.4 meters, and the distance between the laptop and the wireless router is about 0.2 meters.

7.2 Data collection

To evaluate the performance of our system, we collected training and testing data from 3 users. These three users are university students whose heights and sizes are quite different. Moreover, none of them have knowledge of Wi-Fi-based human recognition systems. For each sign language, every user will provide 200 instances at the distance of about 0.3 meters.More specially, 100 instances of each gesture are used as training datasets, and the remaining 100 instances of each gesture are used as testing dataset. At the training stage, in order to get the accurate starting timestamp of each gesture, we use an extra timer. We ask the user to follow the timer and perform the gesture every 5 seconds. All the gestures are asked to to be done with same duration of 2 seconds.

7.3 Prediction performance

Fig. 6(a) shows the mean prediction accuracies of our prediction combination model and classification models on each Wi-Fi receiver with threshold $\tau = 0$. We can see that our system can improve recognition performance for all supported gestures. Since "Good bye" is easy to classify on the first receiver, both of the first receiver and our system have great prediction accuracy of 100%, while the second receiver only has prediction accuracy of 88%. For "Hello", the prediction accuracies are 73% and 75% on the first and second receiver, while our system has better result of 90%. For "You're
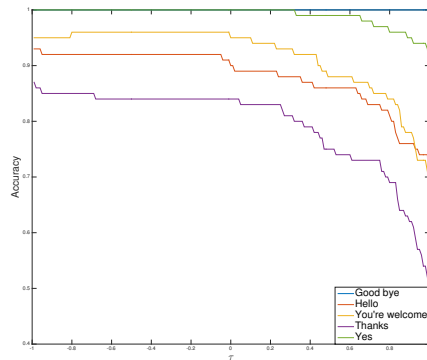
Figure 7: Influence of different τ

welcome", we improve the accuracy on the first receiver from 45% to 95%. For "Thanks", the accuracy is improved by 24% and 8%, respectively. Similarly, the accuracy of "Yes" is improved by 19%. The experiment results show that we can get more accurate sign language estimation with multiple CSI sources by weighted majority voting.

7.4 False positive

In addition to the prediction accuracy, we further explore the false postive of each gesture supported. We use the same dataset that is used in the last subsection, and the evaluation result is illustrated in Fig. 6(b). It is clear that the false positive of prediction can be improved by using two receivers in most cases. For "Goodbye", the false positive is reduced from 20.75% and 8.75% to 2.75%. For "Hello", the false positive is 1.25%, which is much lower than that of the first receiver (5%). The false positives of "You're welcome" and "Thanks" are 1% and 0.75%, respectively, which is also much lower than that on any single receiver. Due to the special distribution of "Yes" on each receiver, the false positive is not improved too much, but 2% is already good enough for a human activity recognition system. With more receivers deployed at different locations, we exprect to get more information of one activity from different views. The false positive is expected to be reduce more with more receivers delployed.

7.5 Mean prediction accuracy vs. different users

In order to make sure our system can work for different users. Even for a same activity, the operation range and speed may not be the same since different people tend to have different habits. In this subsection, we add a new experiment to evaluate the influence generated by different users. We use the trained classification model to further evaluate the data collected from the other two participants, and the results are shown in Fig. 6(c). We can find that the prediction performances are still good even if we do not train a new classification model for these two new participants. Although the mean prediction accuracies decrease by about 3.6% and 4.2% respectively, the accuracies are still above 89%. The second user has the worst performance because his habbit is a little different from

the other two users, which in turn lead to the distributions of his instances are different from the first user on the feature hyperplane.

7.6 Prediction accuracy vs. different τ

We further evaluate the influence of choosing different τ during the prediction combination, and the results are shown in Fig. 7. We can observe that the accuracies of most gestures decrease with the τ increases. Since "Goodbye" is pretty easy to distinguish on the first receiver, its accuracy is not influenced by τ. We can also see that "Thanks" is easiest to be influenced by different τ. This is because most instances of "Thanks" are close to the classification boundary and cannot be classified properly on both classifers. When we increase τ, most useful information for recognizing "Thanks" will be discarded, which leads to low prediction accuracy. In our testbed, we set $\tau = 0$ to get good predcition accuracies without introducing to many uncertain predictions.

8 CONCLUSION AND FUTURE WORK

In this paper, we propose a Wi-Fi signal-based indoor sign language recognition system. Different from other existing work, we use two receivers in our system to improve the recognition performance. In the classification stage, different from getting the final prediction from one multi-class classifier, the final recognition result is determined by combining all prediction results on all Wi-Fi receivers. We treat ten classifiers as players and involve them in a weighted voting game. Our experimental results show that our system can get a better mean false positive of 1.55%, and improve the recognition accuracy to 93.8% compared with originial implementation that uses only one laptop in the same environment.

9 ACKNOWLEDGMENT

REFERENCES

[1] *http://dhalperi.github.io/linux-80211n-csitool/*.
[2] *http://www.castlerockfamilyenrichmentcenter.com/works hops/baby-signing/*.
[3] H. Abdelnasser, M. Youssef, and K. A. Harras. Wigest: A ubiquitous wifi-based gesture recognition system. In *Proceedings of the INFOCOM*, pages 1472–1480. IEEE, 2015.
[4] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller. 3d tracking via body radio reflections. In *Proceedings of the NSDI*, pages 317–329, 2014.
[5] K. Ali, A. X. Liu, W. Wang, and M. Shahzad. Keystroke recognition using wifi signals. In *Proceedings of the MobiCom*, pages 90–102. ACM, 2015.
[6] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti. Spotfi: Decimeter level localization using wifi. In *Proceedings of the SIGCOMM*, pages 269–282. ACM, 2015.
[7] Q. Pu, S. Gupta, S. Gollakota, and S. Patel. Whole-home gesture recognition using wireless signals. In *Proceedings of the MobiCom*, pages 27–38. ACM, 2013.
[8] 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*, pages 77–89. ACM, 2015.
[9] 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*, pages 593–604. ACM, 2014.
[10] 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*, pages 65–76. ACM, 2015.
[11] 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*, pages 361–369. IEEE, 2014.