Towards fully domain-independent gesture recognition using COTS WiFi device

Yuqing Yin, Zixin Zhang, Xu Yang, Faren Yan, and Qiang Niu
School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China
✉ Email: niuqiangkd@163.com

This letter proposes a fully domain-independent WiFi-based gesture recognition system based on the multi-label adversarial network. Unlike pioneer machine learning works, our system does not require retraining in the target domain, which benefits from our key idea of eliminating fully domain information such as orientations, locations, environment information. The system proposed by us includes three parts: Feature extractor, domain discriminator, and gesture recogniser. The feature extractor attempts to deceive the domain discriminator based on the adversarial network structure so that it is difficult to judge the input domain label, thereby obtaining domain-independent features and realising fully domain-independent gesture recognition. Extensive experiments are conducted on the Widar 3.0 dataset and our dataset to evaluate system performance. The results show that our system can achieve an in-domain recognition accuracy of 93.2% and a cross-domain recognition accuracy of 87.1%, which is superior to other state-of-the-art works.

Introduction: Gesture recognition is the current research hotspot in the field of human-computer interaction, which has a wide range of applications such as remote control, virtual vision, and security monitoring. Characterised by contact-free, low-cost and privacy-unintrusive, commercial WiFi is wild used for gesture recognition systems, such as E-eyes [1], CSI-based human activity recognition and monitoring system (CARM) [2], WiGest [3], CrossSense [4]. These approaches generally extract statistical or physical features from wireless signals to classify gestures. However, these features contain domain information such as orientations, locations, and environment information. Recently, the Widar 3.0 [5] proposes a zero-effort cross-domain gesture recognition solution, which is a breakthrough work with excellent performance. Besides, it provides a large-scale cross-domain gesture recognition dataset to promote research on gesture recognition. However, the calculation complexity of the Widar 3.0 is very high. Therefore, this solution is difficult to apply to a real-time system.

To address this challenge, we propose a fully domain-independent gesture recognition system based on the multi-label adversarial network. We aim to remove full domain information such as orientation, locations, and environmental information, so that our system does not require retraining in the target domain. Our work is inspired by EI framework proposed in [6], which is designed for identifying cross-environmental human activities such as walking, standing up, and sitting down. However, the EI framework only works well on removing coarse-grained domain information like environment diversity but not involving other factors. Nevertheless, gesture recognition is aimed to identify complex fine-grained gestures that involve multiple body parts. Under the affected environment, location, and orientation of the person, the same gesture will produce different signal results. Therefore, we design a novel multi-label classification model for domain discriminator to further remove multiple domain information. We use softmax as activation of the last layer and binary cross-entropy as a loss function. As a result, domain discriminator can help feature extractors eliminate more domain information and make gesture recogniser better. Besides, a domain-independent feature extractor is the key part of our system, which extracts domain-independent features from the Doppler frequency shift (DFS) profile. Meanwhile, considering that DFS profiles cannot provide rich information like images, deep networks are not suitable for channel state information (CSI) feature extraction. Thus, we design a novel multi-channel shallow network based on convolutional neural network (CNN) for feature extraction. The main contributions of this work can be summarised as follows:

1. We design a novel fully domain-independent gesture recognition system based on the multi-label adversarial network, which can remove the effects of orientation, location, environment without any retraining in the target area.
2. We design a novel multi-channel shallow network based on CNN for CSI feature extraction that can extract domain-independent features from the DFS profile.
3. We conduct extensive experiments on the Widar 3.0 and our datasets to evaluate system performance. The results show that our system can achieve an in-domain recognition accuracy of 93.2% and a cross-domain recognition accuracy of 87.1%. Furthermore, we implement a prototype system, which demonstrates both the effectiveness and real time of our approach.

System design:

Overview: With different locations and orientations of the person and multipath environments, features of the same gesture may vary significantly and fail to serve successful recognition. For example, a person is asked to make the same gesture (push-pull) on four corners of a 2 × 2 m square located at the centre of a room, and the signal features are shown in Figure 1. Besides, a person is asked to make the same gesture (push-pull) at the centre of a room facing different orientations (rear, left, right and front), and the signal features are shown in Figure 2. We can find that location and orientation affect the signal features. Thus, our system is designed to remove multiple domain information.

Our system includes the data preprocessing module and multi-label adversarial learning module as shown in Figure 3. (1) Data preprocessing module: To remove data noise, we use a Butterworth filter and principal component analysis (PCA) for data preprocessing. Then, the short-time Fourier transform (STFT) is utilised to build DFS profiles from CSI data. (2) Multi-label adversarial learning module: We construct a multi-label adversarial learning framework including feature extractor, domain discriminator, and gesture recogniser to remove all domain information. These three parts are cooperatively trained through the adversarial training method. First, we use a batch of data to train the domain discriminator. Since our final goal is to extract a feature containing no domain information and recognise the gesture, the feature extractor and gesture recogniser are trained to deceive the domain recogniser to make it difficult to judge the domain to which the feature belongs, so that the feature extractor can obtain the domain-independent features. To remove multiple domain information, we design a multi-label method in the domain recogniser. Unlike the EI framework, we can remove multi-domain information such as orientation, location, and environment information.

Data preprocessing module: To get DFS profiles, we select two antennas and conjugate multiply their CSI data. Then, the Butterworth filter and PCA method are utilised to filter out the static and high-frequency components, where the filter order is 6 and the cutoff frequency of the filter is 0.25 Hz. Next, we use STFT to transform the CSI data from the time domain to the frequency domain, where the sampling rate is
Gesture recogniser: A classifier is built for gesture recognition based on different gestures. After the feature extractor, we can obtain the feature representation from consecutive convolutional layers to compose a feature extractor. Through a convolution-normalisation-pooling structure twice and construct three features, the DFS profile has less information than image. Specifically, we use the feature extractor to extract domain-independent features and enhance the classification performance of the gesture recogniser. For the domain recogniser itself, it aims to identify the domain where the gesture is performed. To achieve this, the domain discriminator connects the output of the feature extractor $Q$ and the output of the gesture recogniser $\hat{y}$ in series by concatenation operation:

$$F = Q \oplus \hat{y}$$  \hspace{1cm} (2)

Then, we input $F$ into three fully connected layers, like gesture recogniser. Different from gesture recogniser, there are many domain factors such as the environment, location, and orientation of the gesture performance. To deal with the multi-label problem of domain factors, we use sigmoid as the activation function of the last layer to map the output of neurons to a value between 0 and 1 with the sum of them unequal to 1. Then we use the binary cross-entropy loss function to optimise the domain discriminator:

$$\text{Loss}_d (s_i, \hat{s}_i) = -\frac{1}{|D|} \sum_{i=1}^{|D|} \sum_{h=1}^{X_i} y_{ih} \log(\hat{y}_{ih}) + (1 - y_{ih}) \log(1 - \hat{y}_{ih})$$  \hspace{1cm} (3)

where $|D|$ is the number of domain categories and $s_i$ is the domain probability vector of the $i$th input data. The training process of domain recogniser is to minimise the value of $\text{Loss}_d$ and maximise the accuracy of the prediction result. However, it conflicts with our final purpose, that is, to extract a domain-independent feature and to improve the performance of gesture recognition in a new domain. As a result, we design the loss function of our whole adversarial network as below:

$$\text{Loss} = \text{Loss}_d - \gamma \text{Loss}_{\gamma}$$  \hspace{1cm} (4)

where $\gamma$ is the weight parameter, which can try its best to maximise the value $\text{Loss}_d$, and to minimise the value of $\text{Loss}_{\gamma}$, finally making feature extractor fool domain discriminator to get the domain-independent feature and improving the performance of gesture recogniser.

**Experiments:**

**Experiment setting:** We utilise the Widar 3.0 dataset [5] to evaluate our system performance. The Widar 3.0 dataset consists of the CSI, DFS and BVP data of 260,000 gestures under 75 different domains, and it was collected in five locations and five orientations by six receivers. In addition, our prototype system consists of one Wi-Fi transceiver pair as shown in Figure 4. The transmitter is a mini-PC equipped with a cheap Intel 5300 Wi-Fi card working on the 5.24 GHz band, and the receiver is equipped with three antennas. The transmission rate is 500 packets per second. We also build our gesture dataset, named cross-domain dataset, by collecting different gesture data at five positions and four directions in two scenes (living room and bedroom), wherein the total is 6000 gesture data.

**Overall performance:** In this section, we conduct experiments to evaluate our system’s overall performance with no cross-domain test. We

---

500 Hz. Finally, we downsample the DFS profiles to a uniform shape as the model input. If there are $N$ receivers, we can achieve $N$ DFS profiles and stack them into a three-dimension DFS matrix.

**Multi-label adversarial learning module:**

**Feature extractor:** A DFS profile can be regarded as a channel of an image, and DFS profiles constructed by different transceiver devices can be regarded as different colour channels of an image. Therefore, we can reuse the image feature extraction method based on deep learning to construct a feature extractor. We design a novel multi-channel shallow network based on CNN to overcome overfitting problems in the model, as the DFS profile has less information than image. Specifically, we use the convolution-normalisation-pooling structure twice and construct three consecutive convolutional layers to compose a feature extractor. Through the feature extractor, we can obtain the feature representation $Q$ through different gestures.

**Gesture recogniser:** A classifier is built for gesture recognition based on $Q$. Our classifier is a neural network including two fully connected layers to obtain the prediction result $\hat{y}_i$ of $i$ gestures. Besides, we add a dropout layer in the middle of the two layers to address the overfitting problem. To train this model, we use categorical cross-entropy as the loss function to compute the loss between the prediction result $\hat{y}_i$ and the true label $y_i$:

$$\text{Loss}_\gamma (y_i, \hat{y}_i) = -\frac{1}{|X_i|} \sum_{i=1}^{|X_i|} \sum_{r=1}^{|Y_i|} y_{ir} \log(\hat{y}_{ir})$$  \hspace{1cm} (1)

where $|X|$ is the batch size that refers to the number of training examples utilised in one iteration, and $|Y|$ is the number of gestures.

**Domain discriminator:** Domain discriminator is the key to cross-domain gesture recognition in our system, and its framework is inspired by the EI framework [6]. EI proposes an environment-independent activity recognition, which can only remove environmental information without considering other domain factors. Therefore, we design a novel domain discriminator for removing all domain information to recognise gestures. Cross-domain gesture recognition is a more difficult problem than the recognition task in the fixed domain. The reason is that the data in the new domain needs to be recognised without additional human intervention to achieve cross-domain recognition. As mentioned above, in a different domain, the CSI or DFS profiles of the same gesture have very different patterns, which makes traditional classification methods difficult to accurately recognise the gesture in a new domain. To achieve cross-domain recognition, the feature extractor and gesture recogniser are required to be able to learn the general features of gestures in all domains, which needs the help of domain discriminator. The goal of the domain recogniser in the entire adversarial network is to guide the feature extractor to extract domain-independent features and enhance the classification performance of the gesture recogniser.
divide the training set and testing set randomly and the ratio is 9:1. On the Widar 3.0 dataset and the crossdomain dataset, the average accuracy of our system can reach 93.2% and 90.3%, respectively. The accuracy of the crossdomain dataset is slightly lower than that of the Widar 3.0 dataset because our dataset is collected using only a pair of transceiver devices, and the Widar 3.0 dataset uses six transceiver devices. Figures 5 and 6 are the accuracy confusion matrix of each gesture on the Widar 3.0 and crossdomain datasets, respectively. Overall, our system can perform very well on different gesture datasets.

Cross-domain evaluation: In this section, we evaluate the cross-domain performance of our method. First, we conduct leave-one-out cross-validation on the Widar 3.0 dataset with six DFS profiles of each gesture to evaluate the performance of our method under different kinds of domain factors. As shown in Figure 7, among the average accuracy of the six gestures, ‘drawing N’ gesture achieves the best performance of 94.31% when performed at a new location. We can also find that the overall cross-location accuracy can reach 89.7%. For the other domain factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%. We also train our model with crossdomain dataset and test its cross-domain performance using the data of new room, location, and orientation factors, our method can achieve the average cross-orientation accuracy of 81.3%

Method comparison: In this section, we compare with other state-of-the-art wireless sensing works (such as Widar 3.0 [5], CARM [2], EI [6] and CrossSense [4]) in performances of both in-domain and crossdomain recognitions on the Widar 3.0 dataset. CARM utilises the DFS profile as a feature and hidden Markov model as a classifier. EI uses adversarial networks, but it does not consider multiple domain labels but only removes environmental information. CrossSense proposes an artificial neural network-based model to achieve the signal feature translation between source and target domains. The comparison results are shown in Figure 8, and we can find that our work achieves the best results in overall accuracy. The accuracy of the Widar 3.0 is relatively close to our work, but the time to calculate the speed spectrum requires a relatively long calculation time even under the condition of parallel calculation. It takes about 150 s to calculate the BVP from the Doppler spectrum. Our system’s running speed is about 45 times that of the Widar 3.0, so our system may have better real-time performance.

Conclusion: This letter proposes a fully domain-independent WiFi-based gesture recognition system based on the multi-label adversarial network. This system can remove fully domain information such as orientations, locations, environment information, and so on. We conduct an extensive experiment on the Widar 3.0 dataset and a designed prototype system. The results show that our system has a better performance than the current state-of-the-art gesture recognition system using the COTS WiFi device.

Acknowledgements: This work has been supported by The National Natural Science Foundation of China (51774282). The Widar 3.0 dataset can be available from http://tns.thss.tsinghua.edu.cn/widar3.0/, while crossdomain dataset is available on request from the authors.

References
1. Wang, Y., et al.: E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures. In: Proceedings of the 34th Annual International Conference on Mobile Computing and Networking, Maui, Hawaii, USA, pp. 617–628 (2018)
2. Wang, W., et al.: Device-free human activity recognition using commercial WiFi devices. IEEE J. Sel. Areas Commun. 35(5), 1118–1131 (2017)
3. Abdelnasser, H., Youssef, M., Wigest, HK.A.: A ubiquitous WiFi-based gesture recognition system. In: 2015 IEEE Conference on Computer Communications (INFOCOM), Kowloon, Hong Kong, pp. 1472–1480 (2015)
4. Zhang, J., et al. CrossSense: Towards cross-site and large-scale WiFi sensing. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, pp. 305–320 (2018)
5. Zheng, Y., et al. Zero-effort cross-domain gesture recognition with Wi-Fi. In: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, Seoul, Korea, pp. 313–325 (2019)
6. Jiang, W., et al. Towards environment independent device free human activity recognition. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India pp. 289–304 (2018)