Device-Free Gesture Recognition Using Time Series RFID Signals

Han Ding[1](✉), Lei Guo[2], Cui Zhao[1], Xiao Li[2], Wei Shi[1], and Jizhong Zhao[1]

[1] School of Computer Science and Technology, Xi'an Jiaotong University,
Xi'an, China
{dinghan,zjz}@xjtu.edu.cn, zhaocui@stu.xjtu.edu.cn, weishi0103@sina.com

[2] School of Software and Engineering, Xi'an Jiaotong University, Xi'an, China
{gl0103,lixiao0906}@stu.xjtu.edu.cn

Abstract. A wide range of applications can benefit from the human motion recognition techniques that utilize the fluctuation of time series wireless signals to infer human gestures. Among which, device-free gesture recognition becomes more attractive because it does not need human to carry or wear sensing devices. Existing device-free solutions, though yielding good performance, require heavy crafting on data preprocessing and feature extraction. In this paper, we propose RF-Mnet, a deep-learning based device-free gesture recognition framework, which explores the possibility of directly utilizing time series RFID tag signal to recognize static and dynamic gestures. We conduct extensive experiments in three different environments. The results demonstrate the superior effectiveness of the proposed RF-Mnet framework.

Keywords: Gesture recognition · RFID · Device free

1 Introduction

Human gesture recognition, which enables gesture-based Human-Computer Interaction (HCI), plays an important role in a wide range of applications, such as smart home, health care, especially the support for sign language (for example American Sign Language (ASL)) which can benefit the life of people who are deaf or hard of hearing. Traditional smart devices, say smart phones/pads, watches, and other wearable sensors, are widely used to recognize human gestures. However, such device-based approaches have the limitations that users need to carry or wear the devices, which are not convenient and feasible to real-world applications.

To overcome above limitations, a lot of effort have been made to explore device-free human gesture recognition techniques. Such methods usually require

This work is supported by NSFC Grants No. 61802299, 61772413, 61672424, Project funded by China Postdoctoral Science Foundation No. 2018M643663.

Q. Li et al. (Eds.): BROADNETS 2019, LNICST 303, pp. 144–155, 2019.
https://doi.org/10.1007/978-3-030-36442-7_10

users to perform hand or body motion and recognize these motions to be the HCI operations. One possible approach is to use imagery-based devices, *e.g.*, Kinect and LeapMotion [2,3], to track human motions in a natural way. However, these systems require line-of-sight and raise concerns on user privacy. Another approach, which is our focus in this paper, is to use radio frequency (RF) based sensing techniques of wireless devices. Generally, such wireless signal based solutions identify the gesture based on the rationale that there will be changes (*i.e.*, amplitude or phase) of time series RF signals when human performs specific gesture before the sensing system. By extracting certain features [7,8], the system can recognize the gesture through pre-defined similarity measures, such as Dynamic Time Warping (DTW) or distance-based classifiers. However these approaches all need heavy crafting on data preprocessing and feature engineering. And the performance is highly dependent on the selection of feature extraction algorithms.

In recent years, convolutional neural networks (CNN) have led to impressive success on objection recognition, audio classification, *etc.* [12,13]. A key superiority of CNN is its ability to automatically learn complex feature representations using its convolutional layers. Inspired by this, it is natural to ask a question: is it possible to automatically learn the feature representation from time series and realize the gesture recognition? In this paper, we design a multi-branch CNN network, namely RF-Mnet, for profiling time series and classifying gestures.

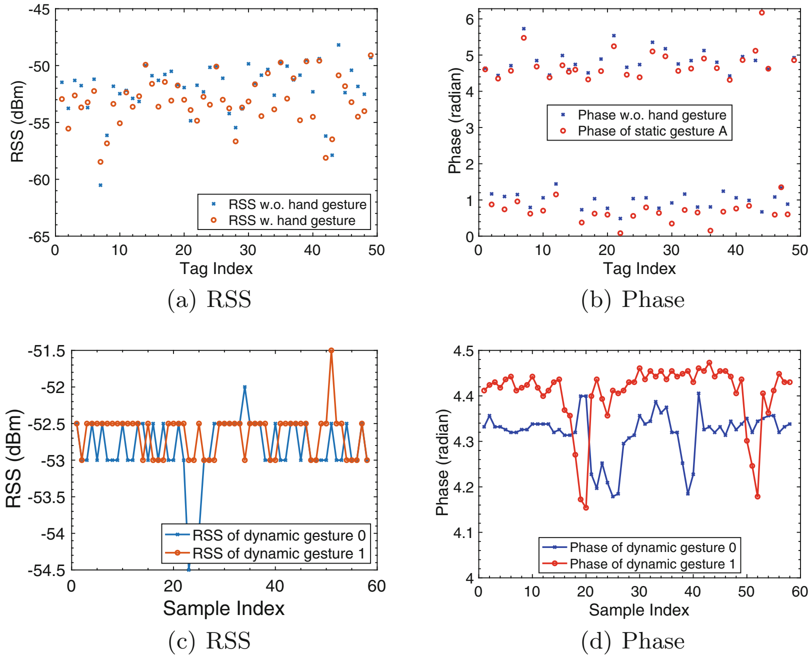
Specifically, with the rapid development of RFID techniques, RFID tag is no longer only an identity of certain product, it serves as wireless sensor for various applications [10,24,27]. The passive tag has the property of low cost and battery-free access. Inspired by this, we explore the possibility of recognizing human gestures via RFID systems. Our prototype of RF-Mnet is shown in Fig. 3(a). We deploy an array of 49 passive tags (7 × 7 tag array) as the sensing plane. RF-Mnet do not need the user to carry any devices. The user just performs the gestures (including static gestures and dynamic gestures) in the air before the plane. The induced variations of RF signals (*i.e.*, RSS and phase) can be collected and correlated to the gestures. We implement a prototype of RF-Mnet using Commercial-off-the-shelf (COTS) RFID devices, and extensive experiments prove the effectiveness of our solution.

2 RF Signal Properties

Our RF-based motion estimation relies on transmitting the RF signal and receiving its replies (*i.e.*, reflections). In our system, we adopt the widely-used wireless technology: UHF RFID.

To capture the reflections from human, we deploy an RFID tag array. The human body is made up primarily of water from the RF point of view. Water is strongly absorptive around 900 MHz (with the dielectric constant of around 80 at room temperature), and radio waves are reflected by body parts. That is, the human body is both a reflector and an absorber of RF energy.

In ideal circumstances (like the *anechoic chamber*), the RF wave leaves from the reader antenna and strikes the tag. However, in real scenarios in which most



(a) RSS (b) Phase

(c) RSS (d) Phase

Fig. 1. RSS and phase distribution (a)(b) w./w.o. hand gesture. (c)(d) dynamic gesture 0 and 1.

RFID systems are used, the wave emitted from the reader antenna will interact with many other objects besides the tag itself. For example, in a typical office building, the integrated backscattered signal of a tag shall be the addition of the *direct* beam along the path between the reader and the tag, and those that are *reflected* (*i.e.*, from the floor, a distant wall, and nearby furniture). We can write the resulting signal as:

$$S_{total} = S_{dir} + S_{ref_0} \tag{1}$$

Similarly, let us consider, when the human body (*i.e.*, wonderful reflector) exists, the newly resulting signal of a tag is the interaction of S_{total} and the reflection (S_{ref_h}) from the human:

$$S_{\hat{total}} = S_{total} + S_{ref_h} \tag{2}$$

Human reflected wave (S_{ref_h}) will add to (*i.e.*, in phase) or subtract from (*i.e.*, out of phase) S_{total}, causing the received signal vary. Specifically, this interaction happens even when the human body is far from the direct beam from a tag.

Preliminary Experiment: Typical commercial off-the-shelf (COTS) RFID reader (*e.g.*, Imping R420) can report the channel parameters, *i.e.*, received signal strength (RSS) and phase, of each interrogated tag. To investigate the influence of human to backscattered tag signals, we conduct a group of proof-of-concept experiments. We first collect and calculate the average RSS and phase value of each tag in a 7×7 tag array in the static environment. Then the volunteer performs a static hand gesture (*i.e.*, letter 'A' as shown in Fig. 3(b)) before the tag array. Figure 1(a) and (b) compare the RSS and phase of each tag with and without the hand gesture. We can observe that almost all tag signals vary (*e.g.*, increase or decrease) with a human hand nearby the array, which demonstrates that the reflected wave from human body has essential importance. In addition, we also let the volunteer perform two dynamic gestures, *e.g.*, moving the hand to write the number 0 and 1 in the air. Figure 1(c) and (d) illustrate the RSS and phase of tag #1, which tell that different gestures have different waveform profiles. In a nutshell, the received signals of tags contain the information of human reflections, inspiring us to infer the human gesture using RF time series signals.

3 Method

In this section, we define the RFID time series classification problem. Then we introduce the RF-Mnet framework.

3.1 Problem Definition

A time series is a sequence of data points with timestamps. In this paper, we use 7×7 tags in the implementation. We denote the time series of tag i as $T_i = t_{i1}, t_{i2}, \ldots, t_{in}$, where t_{ij} is the value at timestamp j and there are n timestamps. Thus, a real time series of RF-Mnet is

$$\mathcal{T} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_m \end{bmatrix} = \begin{bmatrix} t_{11}, t_{12}, \ldots, t_{1n} \\ t_{21}, t_{22}, \ldots, t_{2n} \\ \vdots, \vdots, \cdots, \vdots \\ t_{m1}, t_{m2}, \ldots, t_{mn} \end{bmatrix} \tag{3}$$

where m is the number of tags ($m = 49$).

A labeled time series dataset is denoted as $D = (\mathcal{T}^k, y^k)_{k=1}^{N}$, which contains N time series and y^k is the associated label. y^k is a real value and $y^k \in [1, C]$, where C is the number of distinguishing labels (*i.e.*, classes). Thus the problem we solve in this paper is to establish a model that can predict an unlabeled time series $\mathcal{T}^k$.

3.2 RF-Mnet Framework

The overall architecture is illustrated in Fig. 2. The RF-Mnet framework has three stages: multi-branch input stage, feature extraction stage, and gesture

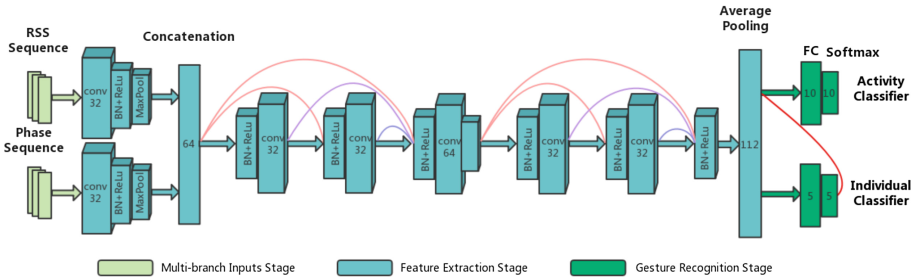


Fig. 2. Overall architecture of RF-Mnet.

recognition stage, in which the input is the time series data, and the output is its label.

Multi-branch Inputs Stage: Different gestures might have different influences on RF channels, such as the change of energy attenuation and propagation path, which are reflected in RSS and phase variations of RF signals. Thus, we take RSS series and phase series as multi-branch inputs, which will provide us a bigger picture of the human motion. Each time series in the multi-branch has the same length.

Feature Extraction Stage: We employ two steps for feature extraction: local convolution and global convolution. The multi-branch inputs ($\boldsymbol{X}^0$) are first fed into the local convolution block which includes a batch normalization (BN) layer, a 2-D convolutional layer (Conv) and a rectified linear unit (ReLu), then the output feature map of local convolution is:

$$\boldsymbol{X} = BN(ReLu(\boldsymbol{W}\boldsymbol{X}^0 + \boldsymbol{b})) \tag{4}$$

where $\boldsymbol{W}$ represents the convolution filters and $\boldsymbol{b}$ is the bias. In particular, the filter size of the Conv layer is 3. The number of filters for both Conv layers is 32. The output of two local convolutional layers will capture a different dimension of features from original signals.

After extracting feature maps from each branch, we then concatenate all features and feed them into the global convolutional stage. Deep convolutional neural networks are proved to be capable of capturing the hierarchy of features [25], where the lower layers respond to primitive features, and the higher layers extract more complex feature informations. Such low and high-level features are both important and complementary in estimating human gestures, which motivates us to incorporate multi-layer information together. Hence, we employ a two-layer DenseNet [12] architecture in global convolutional layers, as shown in Fig. 2. Each layer is a stack of two dense blocks. Each dense block is constructed with two basic blocks. In each basic block, the input of l-th layer is the concatenation of the feature maps produced in all preceding layers $0, 1 \ldots, l-1$. If we denote the sequential operations of BN, ReLU, and Conv as H, the feature map of l-th basic block as $\boldsymbol{X}_l$, then $\boldsymbol{X}_l$ can be calculated is:

$$\boldsymbol{X}_l = H([\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{l-1}]) \tag{5}$$

Between two dense blocks, there is a transition layer which is composed of a BN layer, a ReLU layer, and an 1 × 1 Conv layer followed by a 3 × 3 average pooling layer. The transition layer reduces network parameters by converting the numbers of filter to half.

DenseNet achieves better performance by mitigating vanishing-gradient and enhancing the delivery of features. However, it is possible that the net will overfit the training data since time series data always lacks of complex structures compared with 3-D images that DenseNet is proved to be effective for object detection tasks. Hence, a global average pooling layer (kernel size of 3) is adopted to minimize overfitting and reduce the parameters.

Gesture Recognition Stage: During the gesture recognition stage, we propose two tasks which involve gesture classification and individual classification. The latter is able to authenticate the user identity when s/he conducts a gesture, which can be applied in applications which have privacy and security concerns. The input of individual classifier is the outputs of feature extraction stage ($\hat{\boldsymbol{X}}_{\boldsymbol{i}}$), we then feed the features into a fully connected layer followed by a Softmax activation function. The output of individual classifier can be denoted as:

$$\hat{\boldsymbol{Y}}_{\boldsymbol{i}} = Softmax(\hat{\boldsymbol{W}}\hat{\boldsymbol{X}}_{\boldsymbol{i}} + \hat{\boldsymbol{b}}) \tag{6}$$

where $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{b}}$ are parameters. The output $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ is the predicted possibility of each label for ith input series data. Since gesture classifier contains individual related features, the input of gesture classifier is the concatenation of the output of feature extraction stage ($\hat{\boldsymbol{X}}_{\boldsymbol{i}}$) and the output of individual classifier Y_i. Then, the output of gesture classifier can be described as:

$$\hat{\boldsymbol{Z}}_{\boldsymbol{i}} = Softmax(\hat{\boldsymbol{W}}[\hat{\boldsymbol{X}}_{\boldsymbol{i}}, \hat{\boldsymbol{Y}}_{\boldsymbol{i}}] + \hat{\boldsymbol{b}}) \tag{7}$$

In addition, to train the network, we use cross entropy function to calculate the loss between predictions and the real labels for gesture classification and individual classification. We define L_y as loss function of gesture classifier, and L_z as loss function of individual classifier.

$$\boldsymbol{L}_{\boldsymbol{y}} = -\sum_{c=1}^{N} \boldsymbol{y}_c \log(\hat{\boldsymbol{y}}_c) \tag{8}$$

$$\boldsymbol{L}_{\boldsymbol{z}} = -\sum_{c=1}^{M} \boldsymbol{y}_c \log(\hat{\boldsymbol{y}}_c) \tag{9}$$

where M, N is the number of gesture and individual classes respectively. Then, the composite loss can be denoted as:

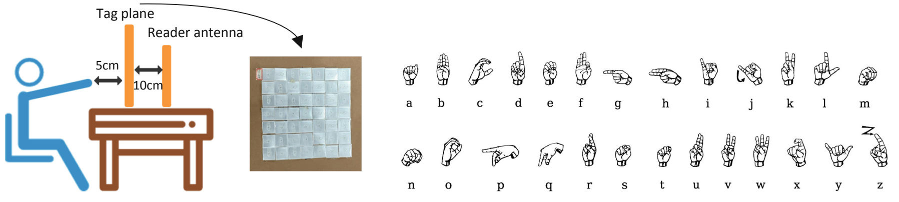
$$\boldsymbol{L} = \alpha * \boldsymbol{L}_{\boldsymbol{y}} + \beta * \boldsymbol{L}_{\boldsymbol{z}} \tag{10}$$

where α, β are hyper-parameters.

4 Implementation and Evaluation

4.1 Implementation

As shown in Fig. 3(a), RF-Mnet consists of an Impinj reader (Speedway R420) and a 7 × 7 Alien-9629 tag array. The whole system runs at the frequency of 922.375 MHz. In order to test the performance of RF-Mnet in multiple environments, we ask 5 volunteers to perform a large number of gestures in three indoor environments: Scene A (tag plane are placed in relatively open space), Scene B (tag plane are placed near walls and tables), Scene C (multiple objects are placed around the RFID tag plane). In each scene, there are people walking around during experiments occasionally.



(a) (b)

Fig. 3. (a) Experimental setups of RF-Mnet. (b) Static gestures: the ASL fingerspelling alphabet [1].

Dataset: In each scene, we collect two kinds of gestures for each person, including static and dynamic gestures. As illustrated in Fig. 3(b), the static gestures are gestures corresponding to 26 English letters specified by American Sign Language (ASL). Dynamic gestures are a handwritten number (0–9,) in the air. The experimental dataset includes 39000 static gestures (5 users × 3 positions × 26 gestures × 100 instances), 15000 dynamic gestures (5 users × 3 positions × 10 gestures × 100 instances).

Parameter Setting: We implement our network with Pytorch 0.4.0. The Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used to train the network. The initial learning rate is 0.001, and it decreases by 50% every 3 epochs. We train the network 40 epochs in total.

4.2 Performance of Gesture Recognition

Overall Accuracy: We first test the overall gesture recognition accuracy of RF-Mnet. In this trail of experiments, 70% and 30% of the data collected in three environments are used for training and testing. The accuracy is shown in

Fig. 4. We can observe that the accuracy of static gesture recognition is higher (say, average accuracy 99.5%). The reason lies in that the human hand is moving during writing the dynamic numbers (*i.e.*, 0–9) in the air. The phase and amplitude of tag signals change dynamically due to the reflections of the moving hand. Thus, dynamic gestures are more susceptible to multipath interference, yielding lower accuracy. However, the average accuracy can still reach 92.5%. The results prove the effectiveness of our framework.

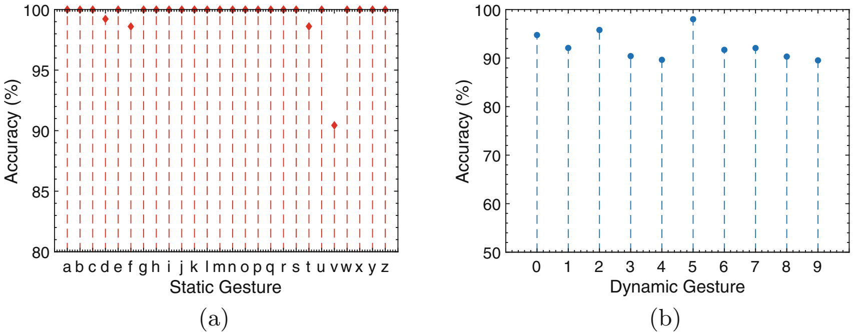


Fig. 4. Overall gesture recognition accuracy. (a) Static gestures. (b) Dynamic gestures.

Impact of Human Diversity: Next, we examine the usability of the system. We invite five volunteers to perform both static and dynamic gestures, 100 times for each. We balance the diversity of the volunteers in terms of their gender (3 males and 2 females), age (ranging from 22 to 28 years old), and other physical conditions (*e.g.*, 158–185 cm in height, 45–70 kg in weight, *etc.*). Note that when performing the dynamic gestures, they are naturally moving their fingers before the tag plane according to their writing habits. Figure 5 compares the average accuracy of static and dynamic gesture recognition. In particular, the accuracy of each volunteer for static gesture recognition is above 95%.

Impact of Distance: We then check the impact of distance between the user hand and the tag plane. We vary the distance from 10 cm to 30 cm. Other settings are consistent as default. We choose four representative static gestures and five dynamic gestures. Specifically, the static gestures involved in this experiment are *a*, *h*, *o*, *v*, and the dynamic gestures are 1, 3, 5, 7 and 9. The average recognition accuracy are plotted in Fig. 6. The average recognition accuracy over five distances is 99.8% for static gestures 90.1% for dynamic gestures. As expected, when enlarging the distance, the accuracy becomes lower. We envision the reason is that a larger distance to may weaken the direct interference from human hand, and involve extra influence from ambient factors, which introduces irregular variations to tag signals and induces lower accuracy.

Impact of Environments: Since the experiments involve three environments, we also compare the accuracy in different scenarios. The accuracy is shown in

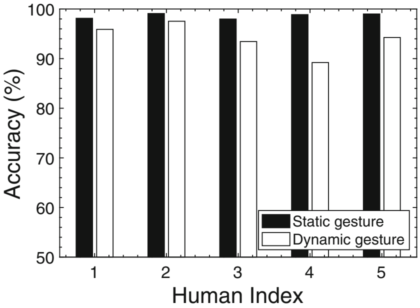


Fig. 5. Accuracy v.s. human diversity.

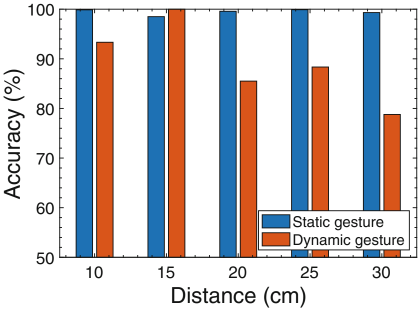


Fig. 6. Accuracy v.s. distance.

Fig. 7. The overall accuracy of three environments reaches 92.4%. In particular, the accuracy of Scene C is lowest because of its rich multipath property.

5 Related Work

Existing studies on the gesture/posture recognition can be classified into following categories:

Computer Vision Based Gesture Recognition: Vision based gesture recognition systems capture fine-grained gesture movements using cameras or light sensors [15,19,20,23,26]. For example, Okuli [26] adopts LED and light sensors to locate user's finger. In-air [19] uses built-in RGB camera of off-the-shelf mobile devices to recognize a wide range of gestures. However, these systems are susceptible to lighting condition changes, which are not suitable for applications where occlusions are everywhere. Most importantly, it will expose user privacy. In contrast, RF-Mnet has no requirement for line-of-sight and is lightweight and scalable.

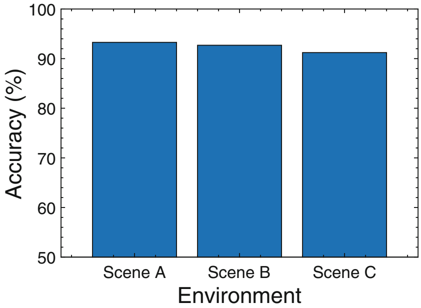


Fig. 7. Accuracy v.s. environment.

RF-Based Gesture Recognition: Prior works on RF-based gesture recognition span a wide spectrum, which can be divided into two categories: device-based and device-free approaches. Device-based methods use wearable devices, such as sensors or tags, for tracking finger or body movements [6,9,11,14,17,18,22]. For example, RF-IDraw [22] leverages beamsteering capability of multiple antennas to detect the direction of the tagged finger and then track the tag by computing the location of intersected beams, which requires a large number of antennas that incurs heavy cost. FitCoach [9] perform fine-grained exercise recognition including exercise types, the number of sets and repetitions by using inertial sensors from wearable devices. These systems rely on wearable devices which are not friendly for users. Another appealing solution is device-free gesture recognition. There exists many systems that track the motion of object by receiving RF signals reflected by objects [4,5,21,24]. For example, WiZ [4] and WiTrack [5] combine frequency modulated continuous wave (FMCW) and multiple antennas technologies for motion tracking. However, both methods require dedicated devices that incur high host for daily gesture monitoring. Tadar [24] arranges a group of tags as an antenna array which receives reflections from surrounding objects and tracks human movements, while it cannot perform fine-grained gesture recognition. Rio [16] detects gestures by touching the surface of the tag with a finger which limits the position of the finger to some extent. In contrast, RF-Mnet is built on COTS RFID devices. Our system enables fine-grained gesture recognition without the need for users to carry the sensing devices.

6 Conclusions

In this paper, we propose an effective deep-learning based framework, namely RF-Mnet, to recognize device-free human gestures. RF-Mnet leverages a COTS RFID tag array as the sensing plane, which allows a user to perform in-air gestures, to capture the time series signals for gesture analysis. Extensive experiments from three environments demonstrate the effectiveness of proposed framework. In particular, RF-Mnet can achieve 99.5% and 92.3% average accuracy for static and dynamic gesture recognition respectively.

References

1. American Sign Language (2019). https://www.nidcd.nih.gov/health/american-sign-language
2. Leap Motion (2017). https://www.vicon.com
3. X-Box Kinect (2017). https://www.xbox.com
4. Adib, F., Kabelac, Z., Katabi, D.: Multi-person motion tracking via RF body reflections (2014)
5. Adib, F., Kabelac, Z., Katabi, D., Miller, R.C.: 3D tracking via body radio reflections. In: Proceedings of USENIX NSDI (2014)
6. Bu, Y., et al.: RF-Dial: an RFID-based 2D human-computer interaction via tag array. In: Proceedings of IEEE INFOCOM (2018)
7. Ding, H., et al.: A platform for free-weight exercise monitoring with RFIDs. IEEE Trans. Mob. Comput. **16**(12), 3279–3293 (2017)
8. Ding, H., et al.: Close-proximity detection for hand approaching using backscatter communication. IEEE Trans. Mob. Comput. **18**(10), 2285–2297 (2019)
9. Guo, X., Liu, J., Chen, Y.: FitCoach: virtual fitness coach empowered by wearable mobile devices. In: Proceedings of IEEE INFOCOM (2017)
10. Han, J., et al.: CBID: a customer behavior identification system using passive tags. IEEE/ACM Trans. Network. **24**(5), 2885–2898 (2016)
11. Hao, T., Xing, G., Zhou, G.: RunBuddy: a smartphone system for running rhythm monitoring. In: Proceedings of ACM UbiComp (2015)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of IEEE CVPR (2017)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of IEEE ICONIP (2012)
14. Mokaya, F., Lucas, R., Noh, H.Y., Zhang, P.: MyoVibe: vibration based wearable muscle activation detection in high mobility exercises. In: Proceedings of ACM UbiComp (2015)
15. Plotz, T., Chen, C., Hammerla, N.Y., Abowd, G.D.: Automatic synchronization of wearable sensors and video-cameras for ground truth annotation-a practical approach. In: Proceedings of IEEE ISWC (2012)
16. Pradhan, S., Chai, E., Sundaresan, K., Qiu, L., Khojastepour, M.A., Rangarajan, S.: RIO: a pervasive RFID-based touch gesture interface. In: Proceedings of ACM MobiCom (2017)
17. Ren, Y., Chen, Y., Chuah, M.C., Yang, J.: Smartphone based user verification leveraging gait recognition for mobile healthcare systems. In: Proceedings of IEEE SECON (2013)
18. Shangguan, L., Zhou, Z., Jamieson, K.: Enabling gesture-based interactions with objects. In: Proceedings of ACM MobiSys (2017)
19. Song, J., et al.: In-air gestures around unmodified mobile devices. In: Proceedings of ACM UIST (2014)
20. Taylor, J., et al.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. ACM Trans. Graph. **35**(4), 143 (2016)
21. Wang, C., et al.: Multi-touch in the air: device-free finger tracking and gesture recognition via COTS RFID. In: Proceedings of IEEE INFOCOM (2018)
22. Wang, J., Vasisht, D., Katabi, D.: RF-IDraw: virtual touch screen in the air using RF signals. In: Proceedings of ACM SIGCOMM (2014)

23. Xiao, R., Harrison, C., Willis, K.D., Poupyrev, I., Hudson, S.E.: Lumitrack: low cost, high precision, high speed tracking with projected M-sequences. In: Proceedings of ACM UIST (2013)
24. Yang, L., Lin, Q., Li, X., Liu, T., Liu, Y.: See through walls with COTS RFID system! In: Proceedings of ACM MobiCom (2015)
25. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
26. Zhang, C., Tabor, J., Zhang, J., Zhang, X.: Extending mobile interaction through near-field visible light sensing. In: Proceedings of ACM Mobicom (2015)
27. Zhao, C., et al.: RF-Mehndi: a fingertip profiled RF identifier. In: Proceedings of IEEE INFOCOM (2019)