2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CTGR: A Dual-Branch Convolutional-Transformer Network for RFID-Based Contactless Gesture Recognition

Ruofan Ma[1], Lvqing Yang[2]✉, Yongrong Wu[2], Qianwen Mao[2], Wensheng Dong[3], Yifan Liu[4], Ziyan Wen[5], Bo Yu[3], and Yishu Qiu[2]

[1] Institute of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen, China
[2] School of Informatics, Xiamen University, Xiamen, China
[3] Zijin Zhixin (Xiamen) Technology Co., Ltd., Xiamen, China
[4] University of California San Diego California, United States
[5] Engineering and IT The University of Melbourne Melbourne, Australia
lqyang@xmu.edu.cn

Abstract. RFID-based gesture recognition, while overcoming vision-based limitations in privacy and environmental robustness, faces three key challenges: inadequate temporal dynamics modeling, disjointed local-global feature integration, and suboptimal fusion of complementary Received Signal Strength Indicator (RSSI) and phase signals. To address these limitations, we present CTGR, a dual-branch CNN-Transformer architecture for RFID gesture recognition. Our CTGR framework first establishes dual parallel input pathways for RSSI and phase signals, then processes each branch through synergistic components: the STC layer employing depthwise separable convolutions to extract noise-robust local features from both modalities, and the MATE module applies multi-head self-attention to capture global temporal dependencies in each signal domain. Finally, CTGR combines the processed RSSI and phase features through a strategic fusion mechanism that effectively integrates their complementary properties, enabling comprehensive modeling of gesture dynamics. Extensive experiments across diverse scenarios validate the method's exceptional effectiveness, achieving a 97.38% average accuracy on a 7-class gesture dataset. Compared with mainstream recognition algorithms, CTGR demonstrates superior robustness in adapting to diverse users, varying gesture speeds, and challenging environmental conditions, ensuring consistent performance across real-world scenarios. This work enhances RFID-based interaction systems through spatio-temporal feature fusion, offering practical solutions for robust human-machine interfaces in dynamic environments.

Keywords: RFID, Gesture Recognition, CNNs, Transformer, Spatio-Temporal Features

1 Introduction

Gesture recognition has become crucial for human-machine interaction (HMI) and augmented reality (AR) applications, spanning smart homes to healthcare [1]. While computer vision achieves high accuracy [2], it is limited by lighting sensitivity and

privacy concerns. Alternatives like WiFi [3], radar [4], and acoustic systems [5] face hardware or environmental limitations. RFID-based solutions [6, 7] demonstrate superior practicality due to their contactless operation, portability, and cost-effectiveness.

However, traditional RFID gesture recognition methods, which often require users to wear tags (such as the RF-glove [8] with five tags per finger), have limitations in practical applications like elderly care. Fortunately, recent work, like RF-Finger [9], has addressed this issue by employing tag arrays for contactless interaction, thus enhancing usability. Meanwhile, the integration of artificial intelligence with RFID has gained traction, with methods like ReActor [10] using Random Forest for feature extraction and Wang et al.[9] applied convolutional neural networks (CNNs) for gesture recognition.

Although these methods demonstrate strong performance in static feature extraction, they still face three critical challenges: First, they exhibit significant limitations in processing temporal-domain information, as they typically focus exclusively on either local or global feature extraction while failing to account for the complex interplay between them. Second, due to these deficiencies in temporal modeling and feature interaction, existing approaches underperform when handling complex and fine-grained gestures (e.g., continuous gestures or multi-finger coordinated motions). Third, existing methods fail to effectively combine RSSI and phase signals, which capture complementary angles but have different noise profiles and spatio-temporal behaviors, resulting in suboptimal fusion for precise gesture recognition.

The complementary strengths of CNNs and Transformers motivate our hybrid design: CNNs efficiently extract local spatio-temporal features through hierarchical convolutions [11], while Transformers capture global dependencies via self-attention [12]. This synergy has proven effective in sequential data processing (e.g., [13, 14]), suggesting its potential for modeling RFID-based gesture sequences, where both fine-grained motions and long-range temporal patterns are critical.

To this end, we propose CTGR, a novel dual-branch CNN-Transformer hybrid network that addresses key challenges in RFID-based gesture recognition through synergistic local-global feature learning. The architecture consists of two core components: (1) a Spatio-Temporal Convolutional Layer (STC) that employs depthwise separable convolutions to extract local motion patterns from RFID signals while suppressing noise, and (2) a Multi-head Attention Encoding Layer (MATE) that captures global temporal dependencies through hierarchical self-attention. The system processes phase and RSSI signals through parallel STC-MATE branches, followed by adaptive feature fusion for final classification. The key achievements of our research are as follows:

1) We propose a dual-branch architecture for RFID signal processing that decouples RSSI and phase feature modeling while enabling adaptive cross-signal fusion, overcoming the limitations of suboptimal signal integration in prior works.

2) A novel STC-MATE hybrid network architecture is designed, in which the STC module focuses on local feature extraction, while the MATE module effectively models global temporal dependencies, achieving a comprehensive characterization of gesture dynamics. This architecture unifies local-global gesture dynamics for the first

time in RFID sensing, resolving the long-standing trade-off between motion granularity and temporal context awareness.

3) Our method achieves 97.38% recognition accuracy across diverse real-world scenarios including user variability, speed fluctuations, and environmental noise, demonstrating significant improvement over existing approaches. These results validate the effectiveness of our hybrid local-global spatio-temporal feature integration framework in addressing key limitations of current gesture recognition systems.

2 Related work

2.1 Gesture Recognition Technologies

Gesture recognition systems have evolved through two key paradigms: contact-based and contactless approaches. Early RFID systems required physical tag attachment through wearable devices (e.g., Magic Ring [15]) or tagged objects (ShopMiner [16], CBid [17]), often combined with external sensors (Kinect [18]) or antenna arrays (RF-Dial [19]). While effective, these contact-based methods impose wearing constraints that hinder deployment in sensitive applications like elderly care. Recent contactless alternatives include vision-based (In-air [20]), acoustic [21]), and millimeter-wave [22]) systems, yet suffer from lighting dependence, range limitations, and privacy concerns. While newer RFID solutions like RF-Finger [9] and GRfid [6] have addressed these limitations, they primarily rely on template matching or shallow machine learning models. These approaches fail to effectively model temporal dynamics or fuse complementary signal modalities (RSSI and phase), leading to reduced accuracy and insufficient robustness in practical scenarios.

2.2 Local-Global Feature Interaction in RFID Sensing

Existing approaches to RFID-based gesture recognition exhibit a dichotomy: CNN-based methods [11] capture local temporal features but neglect global dependencies, while Transformer variants [12] model long-range patterns at the expense of local signal details. Recent CNN-Transformer hybrids have demonstrated remarkable success in domains like computer vision [13] and natural language processing [14]. However, these architectures have not been effectively adapted for RFID-based sensing due to unique challenges including signal noise, spatio-temporal variability, and the need for multi-signal fusion - all of which demand specialized solutions.

3 Preliminaries

In this section, we elaborate on the fundamental theories of RFID technology, analyze the impact of gesture movements on RFID signal propagation, and discuss the arrangement of tag arrays.

3.1 Definition of RFID Signal

RFID-based contactless sensing utilizes tag-reflected signal variations to infer object states in backscatter communication environments. As demonstrated in [8], proximity of conductive objects (e.g., fingers) induces measurable fluctuations in RSSI and phase signals, as shown in Fig. 1(a). The tag's response signal can be modeled as:

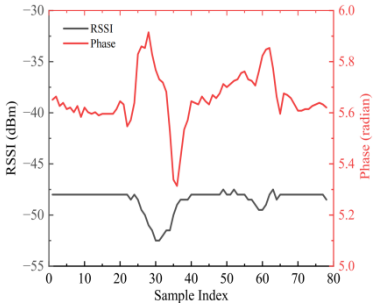
$$S = \sqrt{\frac{10^{\frac{R}{10}}}{1000}}cos\theta + J\sqrt{\frac{10^{\frac{R}{10}}}{1000}}sin\theta, \quad (1)$$

where R is the RSSI value and θ is the phase value.

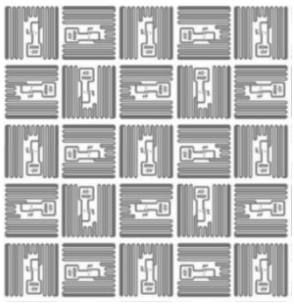
When the finger is placed in front of the tag, the signal returned S consists of two components: the signal S_{direct} transmitted directly to the tag, and the reflected signal $S_{reflect}$ passing through the finger. For S_{direct}, it is directly measurable and equal to the signal the tag sends when the hand is down. Therefore, we can determine the signal reflected off the finger by subtracting S_{direct} from the S computed while the hand is facing the tag array. The impact of finger motions on the RFID signal is likewise represented by this reflected signal.

3.2 Tag Deployment

A dense tag array is essential for capturing gesture-induced signal variations but may cause mutual coupling between adjacent tags. To mitigate this, we adopt an orthogonal arrangement, which minimizes interference and ensures stable signal measurement [23]. The tag array layout, as illustrated in Fig. 1(b), employs an orthogonal placement to minimize mutual coupling between adjacent tags.



(a) Illustrations of original RFID signal.



(b) Tag array layout.

Fig. 1. RFID Signal Characteristics and Tag Array Layout: (a) Signal reflection model under finger gestures; (b) Orthogonal 5×5 tag deployment minimizing mutual coupling.

4 System Design

4.1 Problem Definition

During the RFID signal acquisition process, the collected data is a sequence of data points with timestamps. In this paper, we denote the time-series data received from tag i as $x_i(t)$, which represents the data of tag i at time t. Therefore, the time-series data of a gesture can be expressed as:

$$X = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ x_m(1) & x_m(2) & \dots & x_m(T) \end{bmatrix}, \quad (2)$$

Among them, m represents the number of tags. Since we used a 5×5 tag array, m= 25 in this case. T represents the duration of the action. Therefore, our data set can be expressed as $D = (X_k, Y_k)_{k=1}^{N}$, where N represents the number of data sets, and $Y_k \in [0,6]$ denotes the gesture category ("a","b","c","d","e","d_u","l_r") associated with the data. Here, labels "d_u" and "l_r" correspond to down-to-up and left-to-right hand gestures respectively.

4.2 System framework

The overall architecture of CTGR, depicted in Figure 2, comprises several key components: (1) Preprocessing aligns and filters raw data, (2) HMM-based Segmentation isolates individual gestures, (3) Dual-path Feature Extraction (STC for local patterns, MATE for global dependencies) processes RSSI and phase signals separately, and (4) Attention-based Fusion combines features for final classification. This streamlined architecture achieves robust recognition while maintaining computational efficiency.

Preprocessing. The raw RFID signals undergo three key steps: (1) timestamp alignment via linear interpolation to ensure uniform sampling intervals, (2) phase unwrapping [24] to resolve 0-2π discontinuities caused by cyclic measurements, and (3) noise reduction using a Savitzky-Golay filter [25] that preserves signal morphology while suppressing high-frequency artifacts.

Gesture Segmentation. Precise segmentation of RFID gesture signals is nontrivial due to the inherent tension between noise suppression and motion preservation - incorrect boundaries directly propagate to recognition errors [6]. To address this challenge, we adopt a Hidden Markov Model (HMM)-based segmentation approach that models the joint RSSI-phase feature space using Gaussian hidden Markov models to capture temporal dependencies in gesture signals. This method learns latent state transitions to robustly identify gesture boundaries even under noisy conditions. The HMM framework effectively handles uncertainties in signal patterns while maintaining temporal consistency of segmented gestures.

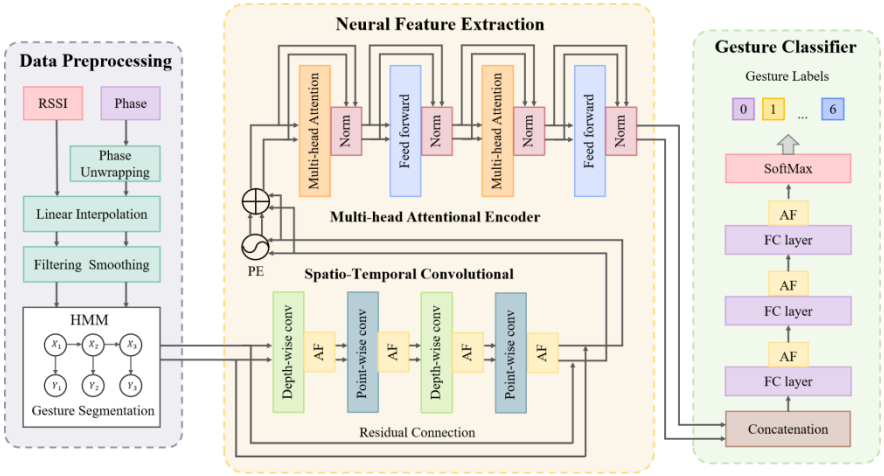


Fig. 2. Architecture of the proposed CTGR model. The proposed system processes RFID signals through sequential preprocessing, segmentation, dual-branch STC/MATE feature extraction, and fused classification.

Multi-branch Inputs. To capture the diverse effects of gestures on RF channels, we incorporate both RSSI and phase signals as multi-branch inputs. This approach allows the model to extract unique information from each signal type, enhancing its ability to recognize complex gestures.

Spatio-Temporal Convolutional Layer (STC). The STC module serves as the primary feature extractor in our system, employing optimized convolutional operations to simultaneously capture local spatiotemporal patterns and attenuate noise in multi-branch RFID time-series data.

The STC module employs an optimized depthwise separable convolution architecture comprising three key components: (1) depthwise convolutions that process each RFID tag's signal as an independent feature channel to extract localized temporal patterns, (2) pointwise convolutions that integrate cross-channel features, and (3) residual connections that enhance gradient flow and model generalization. To preserve temporal causality in gesture sequences, we implement causal 1D convolutions with zero-padding, ensuring the output at each timestep depends solely on preceding inputs. This design effectively isolates gesture-specific spatiotemporal patterns while preventing information leakage and maintaining temporal dependencies across the sequence.

After the deep convolution, the pointwise convolution (1×1) is used to combine the channel features in the spatial dimension. It integrates cross-channel features through learned linear combinations, enabling effective information exchange between different RFID tag channels while preserving spatial relationships. Additionally, appropriate dimension adjustments are made to match the input of the MATE, and finally, a

new feature map X_{new} is generated. The entire process described above can be expressed as:

$$X_{new} = Conv_{Point}(Conv_{Depth}(X_t)), \quad (3)$$

Multi-head Attentional Encoder Layer (MATE). The Multi-head Attentional Encoder Layer (MATE) employs a standard multi-head self-attention mechanism coupled with multi-layer perceptrons (MLPs) to model global temporal dependencies in gesture sequences. By processing parallel attention heads, MATE effectively captures both local gesture details and global temporal patterns, while the MLPs provide nonlinear feature transformations. The architecture incorporates residual connections with batch normalization to ensure stable gradient flow during training, enabling robust learning of complex gesture dynamics. This integrated design allows simultaneous analysis of fine-grained motion features and overarching gesture semantics through its attention-based processing framework.

As a critical component, the Multi-Head Attention (MHA) enables the model to simultaneously attend to the information of the input data within different representational subspaces. This not only strengthens the ability to capture local features but also improves the understanding of the overall trend of gestures. Specifically, the self-attention of the i-th head generates the input matrices Q_i, K_i, V_i, and for self-attention by multiplying the input X_0 with several learnable weight matrices W_Q, W_K and W_V. Subsequently, the attention scores are calculated using the softmax function, and then multiplied by the V_j to obtain the result. The formula is presented as follows:

$$head_i = Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}}) \cdot V_i\,, \quad (4)$$

The multi-head attention outputs are concatenated and linearly transformed through a trainable projection matrix W^0 to synthesize features from different representation subspaces. This design enables comprehensive feature learning by jointly attending to information across multiple attention heads. The formula is as follows:

$$MHA = Concat(head_1, \ldots, head_n)W^0\,, \quad (5)$$

The MLP module implements a nonlinear feature transformation through two fully connected layers: first expanding the input features into a high-dimensional space $R^{\mathrm{d}} \rightarrow R^{\mathrm{k}}$ (where $k > d$) via learnable weights W_1 to enhance feature separability, then compressing back to the original dimension $R^{\mathrm{k}} \rightarrow R^{\mathrm{d}}$ through W_2 to select the most discriminative features. This bidirectional dimensional transformation is formally expressed as:

$$MLP = W_2(Relu\,(W_1X_1 + b_1)) + b_2\,, \quad (6)$$

Here, W_1 and W_2 represent the weights of the two fully connected (FC) layers, $Relu$ is the non-linear activation function, and b_1 and b_2 denote the biases.

The MATE module's output undergoes adaptive average pooling [26] to generate a fixed-length feature vector, enabling robust handling of variable-length gesture sequences while preserving discriminative spatiotemporal patterns for classification.

Gesture Classifier. The outputs from both branches' STC and MATE are fused through concatenation, combining their local features and global contexts into a unified high-dimensional representation. This integrated approach provides a comprehensive feature space for gesture recognition.

The fused features are processed through fully connected layers to generate log-softmax classification probabilities, optimized via cross-entropy loss. The definition formula of the cross-entropy function is as follows:

$$G(p,q) = -\sum_{i=1}^{k} p_i \, log(q_i), \tag{7}$$

where p and q are the real and estimated distributions of the i-th class of gesture, respectively.

5 Performance evaluation

5.1 Implementation

Our experimental setup includes a 5×5 Alien-9629 tag array and an ImpinJ Speedway R420 RFID reader equipped with an S9028PCL directional antenna. The tag arrays were mounted on one surface of a white plastic box. A vertically oriented RFID antenna was positioned on the opposite surface, maintaining a consistent separation of 50 cm from the array. Experimental participants performed finger gestures while maintaining a controlled operating distance of 10-15 cm from the array plane. We tested the system in two environments: an empty office (Env_A) and a laboratory with higher interference (Env_B). Ten participants performed five letters and two shapes at varying speeds, generating a dataset of 1400 samples.

5.2 Experimental Evaluation

In order to demonstrate the prediction performance of the proposed CTGR, a large number of necessary qualitative and quantitative experiments are carried out and analyzed in detail.

Experimental Settings. The proposed CTGR model was implemented in Pytorch framework, Python 3.9, and NVIDIA GeForce RTX 3080 GPU environment. We trained the model by optimizing the cross-entropy loss function by setting the batch size to 32, where the Adam optimization algorithm is applied with a learning rate of 0.002 and weight decay of 1×10^{-8}. And gradient clipping is used to prevent gradient explosion. According to the principle of 70% for training and 30% for testing, the dataset was divided into two parts, and 50 epochs were trained.

Experimental Results. We evaluate CTGR against five representative methods: Random Forest [10], RF-Finger [9], TCN, CNN-LSTM, and BiLSTM. As shown in Table 1, CTGR consistently outperforms all comparison methods in recognition accuracy. The overall accuracy rate reaches 97.38%, with the gesture "d_u" achieving 99.10% accuracy. These results demonstrate that CTGR excels not only in capturing local spatio-temporal features from RFID time-series data but also in modeling global temporal dependencies within gesture sequences, thereby achieving superior recognition performance compared to conventional approaches.

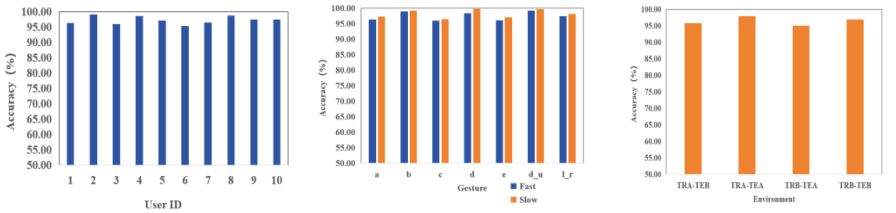
Table 1. Comparison of classification accuracy (%) of different methods. Overall accuracy is computed as macro-average to equally evaluate performance across all gesture categories.

Methods	Overall	a	b	c	d	e	d_u	l_r
Random Forest	76.84	75.32	74.68	79.49	78.35	73.99	77.16	76.75
RF-Finger	88.65	89.26	90.51	93.18	84.68	90.74	89.91	87.01
TCN	92.36	93.48	91.88	94.04	90.21	93.55	95.32	92.23
CNN-LSTM	93.87	93.11	95.99	93.84	89.56	88.98	93.36	92.22
BiLSTM	92.01	91.23	94.12	92.89	91.19	90.89	89.21	93.66
Proposed CTGR	97.38	96.34	98.51	94.92	98.33	96.83	99.10	97.37

Evaluation of Different Users. Model robustness was assessed through ten-subject testing, with the resulting accuracy distribution visualized in Fig. 3(a). The volunteers vary in physical conditions such as gender (six men and four women), age (22 to 28 years old), height (152 to 183 centimeters), and weight (49 to 90 kilograms). The gesture recognition accuracy of all ten volunteers exceeds 95%. Among them, the recognition accuracy of Volunteer 2 (a male) is the highest (98.97%), and that of Volunteer 6 (a female) is the lowest (95.22%). Therefore, the CTGR has relatively good performance in recognizing finger movements of different individuals.

Evaluation of Different Finger Speeds. When performing gesture actions, the volunteers were required to move their hands naturally. However, generally speaking, people cannot always maintain a stable hand movement speed. Therefore, our dataset contains data of different speeds, with half being fast samples and the other half being slow samples. In this part, in order to further evaluate the influence of speed, we conducted a series of new experiments to test the recognition accuracy of CTGR under fingers moving at different speeds, and the results are shown in Fig. 3(b).

Evaluation of Different Environments. We evaluate CTGR's environmental robustness through cross-dataset testing, training on environments A/B (TRA/TRB) and testing on opposite environments B/A (TEB/TEA). The system achieves 95.71% accuracy when trained on TRA and tested on TEB, and 94.97% in the reverse configuration (Fig. 3c). These results validate CTGR's practical applicability in real-world scenarios where environmental stability cannot be guaranteed.



(a) Accuracy with different users.

(b) Accuracy with different gesture speeds.

(c) Accuracy with different environment.

Fig. 3. Evaluation results.

5.3 Ablation study

To rigorously validate each component's contribution, we conduct ablation studies by selectively disabling key modules while keeping other parameters unchanged. The training data, optimization strategy, and evaluation protocol remain identical to the full model setup. As shown in Table 2, the full model achieves 97.38% accuracy, while removing any major module leads to noticeable performance degradation:

STC Module: Removing the Spatio-Temporal Convolutional layer causes a 1.62% accuracy drop (95.76%), confirming its importance for local feature extraction.

MATE Module: The largest performance decrease (-4.17% to 93.21%) occurs when disabling the Multi-head Attention Encoder, highlighting its critical role in modeling global temporal dependencies.

Dual-Branch Input: The 2.8% accuracy reduction (94.58%) demonstrates the necessity of processing RSSI and phase signals separately.

HMM Segmentation: Replacing HMM segmentation with basic thresholding methods degrades accuracy by 3.2%, proving the superiority of our probabilistic boundary detection approach.

Table 2. Performance Ablation Study of CTGR Components.

Modules	Accuracy(%)
CTGR	**97.38**
w/o STC	95.76
w/o MATE	93.21
w/o Dual-Branch Input	94.58
w/o HMM Segmentation	94.18

6 Conclusion

This paper presents CTGR, a novel CNN-Transformer hybrid architecture for RFID-based contactless gesture recognition that addresses three fundamental limitations in

existing approaches: (1) inadequate temporal modeling of gesture dynamics, (2) insufficient integration of local and global features, and (3) suboptimal utilization of complementary RSSI and phase characteristics. Through its innovative dual-branch design featuring dedicated Spatio-Temporal Convolutional (STC) and Multi-head Attention Encoding (MATE) layers, CTGR achieves state-of-the-art performance with 97.38% recognition accuracy while demonstrating robust operation across diverse users, execution speeds, and environmental conditions. The proposed framework not only advances RFID-based gesture interaction systems but also establishes a new paradigm for time-series data processing. Future work will focus on real-time optimization and extension to multimodal sensing scenarios.

Acknowledgments. This work was supported in part by the project "Development of a Cost Intelligent Analysis System for Zijin Mining Equipment" under Project No. 20243160A0723, in part by the project "Research and Application of a Mining Production Control Platform Based on IoT and Big Data"under Serial No. 20240866, funded by the 2024 Industry-University-Research Collaboration Program of the Longyan Xiamen University Integration Research Institute, funded by the Open Research Fund of the State Key Laboratory of Intelligent Manufacturing in Metallurgical Processes (2025).

References

1. Shankar, K. P. K. R., Sawant, E. K., Punarvit, Y.: Arduino based smart gloves for blind, deaf and dumb using sign and home automation mode. In: Proc. Int. Conf. Adv. Technol., Manage. Educ. (ICATME), pp. 244–248. IEEE, Jan. (2021)
2. Swedheetha, C., Naveen, P., Akilan, T., Manikandan, P., Pushpavanam, B.: Computer Vision-Based Hand Gesture Recognition. In: 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), pp. 423–426. IEEE, Gobichettipalayam, India (2024)
3. Zhao, L., Xiao, R., Liu, J., Han, J.: One is Enough: Enabling One-shot Device-free Gesture Recognition with COTS WiFi. In: IEEE INFOCOM 2024 - IEEE Conference on Computer Communications, pp. 1231–1240. IEEE, Vancouver, BC, Canada (2024)
4. Ma, Y., Hui, X., Kan, E.C.: 3d real-time indoor localization via broadband nonlinear backscatter in passive devices with centimeter precision. In:Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, pp. 216–229 (2016)
5. Alemu, M. Y., Lin, Y., Shull, P. B.: EchoGest: Soft Ultrasonic Waveguides Based Sensing Skin for Subject-Independent Hand Gesture Recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering 32, 2366–2375 (2024)
6. Zou, Y., Xiao, J., Han, J., Wu, K., Li, Y., Ni, L. M.: GRfid: A device-free RFID-based gesture recognition system. IEEE Trans. Mobile Comput. 16(2), 381–393 (2017)
7. Liu, Y., Huang, W., Jiang, S., Zhao, B., Wang, S.: TransTM: A device-free method based on time-streaming multiscale transformer for human activity recognition. Def. Technol. 32, 619–628 (2024)
8. Xie, L., Wang, C., Liu, A. X., Sun, J., Lu, S.: Multi-touch in the air: Concurrent micromovement recognition using RF signals. IEEE/ACM Trans. Netw. 26(1), 231–244 (2018)

9. Wang, C., Liu, J., Chen, Y., Liu, H., Xie, L., Wang, W., He, B., Lu, S.: Multi touch in the air: Device-free finger tracking and gesture recognition via cots rfid. In: IEEE INFOCOM 2018-IEEE Conference on Computer Communications. pp. 1691– 1699. IEEE (2018)
10. Zhang, S., Ma, Z., Yang, C., Kui, X., Liu, X., Wang, W., Wang, J., Guo, S.: Real-time and accurate gesture recognition with commercial rfid devices. IEEE Transactions on Mobile Computing (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE (2016)
12. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention Is All You Need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 5998–6008. Curran Associates, Inc., Long Beach (2017)
13. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: CMT: Convolutional neural networks meet vision transformers. arXiv preprint arXiv:2107.06263 (2021)
14. Tay, Y., et al.: Efficient Transformers: A Survey. ACM Computing Surveys (CSUR) 55(6), 1–28 (2022)
15. Jing, L., Cheng, Z., Zhou, Y., Wang, J., Huang, T.: Magic ring: A self-contained gesture input device on finger. In: Proc. 12th Int. Conf. Mobile Ubiquitous Multimedia, pp. 39:1–39:4. ACM (2013)
16. Zhou, Z., Shangguan, L., Zheng, X., Yang, L., Liu, Y.: Design and implementation of an RFID-Based customer shopping behavior mining system. IEEE/ACM Trans. on Netw. 25(4), 2405–2418 (2017)
17. Han, J., et al.: CBID: A customer behavior identification system using passive tags. IEEE/ACM Trans. Netw. 24(5), 2885–2898 (2016)
18. Ren, Z., Yuan, J., Meng, J., Zhang, Z.: Robust part-based hand gesture recognition using kinect sensor. IEEE transactions on multimedia 15(5), 1110– 1120 (2013)
19. Bu, Y., Xie, L., Wang, C., Yang, L., Liu, J., Lu, S.: RF-Dial: An RFID-based 2D human-computer interaction via tag array. In: Proc. IEEE Conf. Comput. Commun., pp. 837–845. IEEE (2018)
20. Song, J., Sörös, G., Pece, F., Fanello, S. R., Izadi, S., Keskin, C., Hilliges, O.: In-air gestures around unmodified mobile devices. In: Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14), pp. 319–329. ACM, New York, NY, USA (2014)
21. Zhang, C., et al.: Soundtrak: Continuous 3D tracking of a finger using active acoustics. In: Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., pp. 1–25. ACM (2017)
22. Lien, J., et al.: Soli: Ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. 35(4), Art. no. 142 (2016)
23. Han, J., Qian, C., Wang, X., Ma, D., Zhao, J., Xi, W., Jiang, Z., Wang, Z.: Twins: Device-free object tracking using passive tags. IEEE/ACM Trans. Netw. 24(3), 1605–1617 (2015)
24. Itoh, K.: Analysis of the phase unwrapping algorithm. Applied optics 21(14), 2470– 2470 (1982)
25. Schafer, R. W.: What is a Savitzky-Golay filter?[lecture notes]. IEEE Signal Process. Mag. 28(4), 111–117 (2011)
26. Bai, S., Kolter, J. Z., Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv preprint arXiv:1803.01271 (2018)