Journal of Computer and Communications, 2025, 13(6), 47-63
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Survey of Human Pose Recognition Based on WiFi Sensing and Neural Network

Shuo Zhang, Xingshuo Han, Ziheng Meng, Chao Wang, Yuanhang Zhang, Yuqian Ma, Zhengjie Wang*

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, China
Email: 173911717@qq.com, 1435995405@qq.com, 2694418651@qq.com, 2372871477@qq.com, 1781016866@qq.com, 2206911173@qq.com, *cieewangzj@163.com

How to cite this paper: Zhang, S., Han, X.S., Meng, Z.H., Wang, C., Zhang, Y.H., Ma, Y.Q. and Wang, Z.J. (2025) A Survey of Human Pose Recognition Based on WiFi Sensing and Neural Network. *Journal of Computer and Communications*, **13**, 47-63.
https://doi.org/10.4236/jcc.2025.136004

Received: April 30, 2025
Accepted: June 16, 2025
Published: June 19, 2025

Abstract

With technological advancements and increasing user demands, human action recognition plays a pivotal role in the field of human-computer interaction. Among various sensing devices, WiFi equipment has gained widespread application due to its universal presence. This paper explores the use of WiFi devices and neural network technology to achieve human pose recognition for multiple persons, significantly advancing intelligent environmental perception and human motion analysis. First, we review typical applications of human pose recognition based on computer vision, millimeter-wave radar, and WiFi devices. Second, a human action recognition system is designed using WiFi devices. Subsequently, data preprocessing is performed, innovatively applying phase denoising techniques such as unwrapping and linear transformation to CSI signals, and fusing time-frequency analysis, filtering, and deep learning models to accurately correct phases. Then, leveraging deep learning and attention mechanisms, the system accurately determines the positions of multiple persons and learns human pose. By utilizing prior knowledge of human anatomy and Transformer networks to optimize joint features, action recognition accuracy is enhanced. Analysis of typical systems demonstrates that human action recognition based on WiFi devices and neural networks achieves high precision in pose recognition for multiple persons. Finally, existing challenges and future research directions are discussed.

Keywords

WiFi Devices, Neural Networks, Human Pose Recognition, Signal Processing, Attention Mechanism, System Design

1. Introduction

Currently, behavior recognition is a hot topic [1] [2]. Three-dimensional human

pose estimation, a core task in computer vision and artificial intelligence, plays a critical role in numerous fields, such as human-computer interaction, virtual reality, medical monitoring, and security surveillance [3]. However, traditional methods primarily rely on RGB cameras or depth sensors, whose performance is significantly limited in privacy-sensitive scenarios (e.g. homes, hospitals) and environments with insufficient lighting or occlusions. For example, camera-based systems not only require line-of-sight transmission, but also are highly susceptible to environmental factors, making large-scale applications challenging.

In recent years, research on human perception using wireless signals, particularly WiFi, has become a hotspot. This approach offers significant advantages such as non-invasiveness, strong penetrability, no line-of-sight requirement, and privacy protection [4], attracting numerous researchers. Among various wireless signal features, Channel State Information (CSI) has become a key tool for human activity recognition and pose estimation due to its ability to capture fine-grained characteristics of signal propagation (e.g. amplitude and phase of multiple subcarriers) [5]. The development of deep learning has brought significant breakthroughs to CSI-based pose estimation. Neural networks such as CNNs and RNNs are widely used to learn complex patterns from CSI data and map them to human poses while effectively capturing spatiotemporal dependencies in the data [6] [7].

Additionally, fusing CSI with multimodal data further enhances the robustness of pose estimation. Some studies use visual data to generate training labels or supplement critical information in complex environments [8]. In terms of application prospects, this technology demonstrates great potential in medical monitoring (e.g. fall detection for the elderly), security, game (sensorless motion capture), and other fields [9].

This comprehensive review systematically investigates the current state of research concerning pose estimation utilizing CSI and neural network architectures. The study provides a detailed synthesis of existing methodologies and compares the merits and limitations of typical approaches across various applications, including data preprocessing techniques, network models, and feature extraction strategies. Furthermore, this paper examines the challenges associated with environmental variability and the practical difficulties. These challenges, stemming from factors such as multipath fading, dynamic environments, and subject-specific variations, impact the generalizability and reliability of CSI-based pose estimation models. Finally, the review identifies and discusses promising avenues for future research, emphasizing the development of novel algorithms and techniques designed to enhance model generalization capabilities and robustness. This work contributes a valuable resource for researchers seeking to navigate the rapidly evolving landscape of wireless sensing and human activity recognition.

2. State of the Art

2.1. Research Status of 3D Human Pose Recognition Based on Computer Vision

In the field of computer vision, researchers focus on estimating 3D human poses

and shapes from video images. Many scholars have proposed 3D human pose recognition models based on computer vision. For example, Lin *et al*. [10] proposed a model for reconstructing 3D human poses and mesh vertices from single images. It uses an attention encoder to jointly model vertex-vertex and vertex-joint interactions while outputting 3D joint coordinates and mesh vertices. Li *et al*. [11] proposed "Hybrik", a hybrid inverse kinematics method connecting 3D pose estimation and body mesh regression, which converts 3D keypoint positions into 3D human meshes, enabling end-to-end training and forming a closed loop between 3D skeletons and parametric models. This solves the alignment problem of model-based methods and the unrealistic human structure issue in keypoint estimation methods. Jin [12] used the ResNeXt-CBAM encoding network and the parametric human model SMPL to reconstruct 3D humans from single-frame RGB images, effectively improving 3D human reconstruction performance and reducing errors. Although these methods have high recognition accuracy, they are greatly affected by factors such as light and occlusion, may involve privacy issues, and cannot penetrate obstacles like walls.

2.2. Research Status of 3D Human Pose Recognition Based on Wireless Millimeter-Wave Radar

Millimeter-wave radar identifies human 3D poses by transmitting millimeter-wave signals and analyzing phase, amplitude, and other information in the received signals. Pioneering work such as RF-capture [13] demonstrated the possibility of using Radio Frequency (RF) signals for human recognition, sparking significant interest in human pose and shape estimation. Subsequent studies like RF-pose [14] and RF-avatar [15] used finer-grained RF signals to construct human poses and meshes. These works show that RF signals contain sufficient information to estimate human poses and overcome many limitations of traditional vision-based methods, such as poor lighting, clothing interference, and privacy issues. Li *et al*. [16] proposed a precise human pose estimation system based on 77GHz millimeter-wave radar, which generates heatmaps from two sets of radar data and uses CNN to convert 2D heatmaps into human poses. Kwon *et al*. [17] developed a hands-free human activity recognition system using millimeter-wave sensors, whose network protects user privacy and can reconstruct the skeleton of the active human body. Xue *et al*. [18] introduced the first real-time 3D human mesh estimation system using commercial portable millimeter-wave devices, innovatively addressing point cloud sparsity with a deep learning framework, representing complex human meshes with few parameters, generating more realistic meshes using prior knowledge, and using Recurrent Neural Networks (RNNs) to handle missing body part points. Gu *et al*. [19] proposed "mmSense", a multi-person detection and recognition framework that uses the unique sensing characteristics of millimeter waves and an LSTM-based classification model to detect and locate multiple persons simultaneously. While millimeter-wave radar-based 3D human pose recognition is unaffected by lighting conditions, it still cannot penetrate walls and requires expen-

sive specialized hardware.

2.3. Research Status of 3D Human Pose Recognition Based on WiFi Devices and Neural Networks

With the rapid development of information technology, home WiFi devices have become widely used. By collecting and processing radio frequency signals from WiFi, we can generate 3D human poses. Compared with data from video, audio, or optical signals, WiFi signals offer unique advantages: they work in low-light environments, can penetrate walls, effectively protect privacy, and require no specialized hardware (e.g. USRP) or radio signals (e.g. FMCW), making them low-cost. Therefore, this technology has enormous application potential.

WiFi technology uses the transmission characteristics of wireless signals to recognize 3D human poses by analyzing CSI information in signals received by receivers [20]. When humans move within a WiFi coverage area, pose changes affect signal propagation characteristics, allowing pose inference. Common CSI collection tools include the Linux 802.11n CSI Tool [21], ESP32 CSI Tool [22], and Atheros CSI Tool [23]. CSI samples from WiFi systems are often affected by noise and interference, including random phase drift and inversion. Some solutions only use CSI amplitude information while ignoring phase [24], compromising information integrity. The PhaseFi system created by Wang *et al*. [25] used unwrapping and linear transformation to denoise phase signals, effectively preserving integrity.

Jiang *et al*. [26] extracted 2D Angle of Arrival (AoA) spectra from WiFi signals to locate different body parts and used deep learning models to establish complex relationships between 2D AoA spectra and 3D skeletons for pose tracking. Ren *et al*. [27] used a CNN and LSTM deep learning model to abstract 3D human poses from 2D AoA, where CNN extracts spatial dynamics (e.g. limb and torso positions), and LSTM models temporal dynamics (e.g. limb and torso trajectories). Han *et al*. [28] proposed a cross-modal meta-learning method based on Model-Agnostic Meta-Learning (MAML) for few-shot human activity recognition using WiFi, addressing data dependency and scalability issues. Wang *et al*. [29]-[31] demonstrated that commercial WiFi can construct 3D human meshes and estimate 2D AoA of signal reflections using multiple transmit and receive antennas on WiFi devices. Ren *et al*. [32] proposed GoPose, a 3D skeleton-based human pose estimation system using reusable WiFi devices in home environments, suitable for predefined activities at fixed locations but capable of through-wall estimation. Most of these methods are limited to single-person pose recognition, with separate data collection and pose judgment, and can only recognize a few predefined poses, failing to achieve real-time arbitrary pose recognition for multiple persons. Ren [33] *et al*. proposed Winect, a skeleton-based human pose tracking system that does not rely on predefined activities and can simultaneously track free-form movements of multiple limbs in real time. However, it only generates 3D human skeletons without judging body shape, limiting its value in gaming modeling and other fields.

In summary, while there have been research achievements in 3D human pose generation using WiFi devices and neural networks, most current results are limited to non-real-time single-person 3D skeleton generation, with no real-time multi-person 3D contour generation. Therefore, WiFi and neural network-based 3D human contour generation systems have significant development space and application potential.

3. Fundamental Architecture of the Pose Recognition System

This section introduces the fundamental architecture of a pose recognition system. We will then provide a detailed analysis of the function and composition of each component, aiming to clarify the overall system structure and functionalities.

3.1. Data Collection

3.1.1. WiFi Signal Collection

A device (e.g. a laptop) equipped with an Intel 5300 network card, three antennas, and running Ubuntu serves as the core receiver. Professional tools are used to analyze the site's electromagnetic environment, select network card frequency bands and channels, and deploy routers at key positions to build a WiFi network. During experiments, volunteers perform diverse activities in simulated scenarios (e.g. offices, classrooms, corridors), including standing, walking, sitting, and waving. The network card collects CSI data in real time at high frequency and stores it.

When humans move within the coverage area of WiFi transmitters and receivers, their bodies affect wireless signal propagation, causing changes in CSI measurements that contain key information about human poses, providing a basis for subsequent pose and contour analysis based on CSI data.

3.1.2. Camera Data Co-Collection

A Kinect 2.0 camera is introduced, with installation parameters determined based on optical imaging principles and scene geometry to achieve wide coverage of the experimental scene. High-frame-rate, high-resolution videos are recorded to accurately capture human appearance, limb dynamics, and other information. Video data is encoded, labeled, and stored in millisecond-level synchronization with CSI data using high-precision timestamps to ensure accurate correspondence. Stored data is filtered and optimized via image quality assessment algorithms to remove low-quality frames (e.g. blurry or noisy frames), improving data usability.

3.2. Data Preprocessing

3.2.1. WiFi Signal Processing

For CSI phase correction and noise processing, filtering algorithms are studied to remove noise, and phase compensation techniques are used for correction. The applicability of phase compensation algorithms in complex indoor environments is analyzed, and a filter is designed based on the formula:

$$y(t) = H(f) * x(t) + n(t)$$

(where $y(t)$ is the received signal, $H(f)$ is the channel frequency response, $x(t)$ is the transmitted signal, and $n(t)$ is noise) to remove noise and improve signal quality.

3.2.2. Data Fusion

Features from WiFi signal processing are fused with visual features from camera video analysis. Specific algorithms are used to correlate and integrate human position information at the feature level (e.g. a deep learning-based feature fusion network takes CSI [8] features from WiFi signals and visual features from videos as input, learns the correlation between them via the network, and fuses features from different modalities to construct a comprehensive dataset reflecting human states [30]). During fusion, dimensionality reduction or expansion is used to match feature dimensions, and attention mechanisms are applied to weight features based on their importance in describing human states, improving fused data quality.

Some studies have explored combining CSI with visual data, where visual data can generate accurate training labels for CSI-based models or provide supplementary information in hybrid systems fusing multi-sensor data. The fusion method enhances pose estimation accuracy and provides important references for achieving more precise real-time 3D pose generation for multiple persons [29].

3.3. Algorithm Design and Selection

After studying geometric models and deep learning algorithms, a hybrid approach combining CNN, LSTM, and Transformer is chosen for pose recognition, as geometric models require precise prior knowledge and are limited in complex scenarios, while deep learning algorithms can automatically learn feature representations. Specifically, CNN first extracts spatial features from processed WiFi signal pose data to obtain preliminary representations; Transformer then uses self-attention mechanisms to further mine long-range dependencies between features and enhance feature expressiveness; finally, LSTM learns temporal variation features of poses to achieve accurate real-time pose recognition for multiple persons.

CNN's powerful spatial feature extraction capability allows it to capture spatial information such as limb and torso positions from processed WiFi signal pose data via sliding convolution kernels, obtaining preliminary human pose representations. **Figure 1** shows CNN achieving preliminary human pose representation.

Transformer's self-attention mechanism effectively captures dependency relationships between features at different positions, deeply processing features extracted by CNN to mine potential correlations and enrich feature connotations.

LSTM excels at processing sequential data and learning temporal variation features of poses. Human movements are continuous processes with obvious time-series characteristics. LSTM uses gating mechanisms to effectively memorize past information and update memory states based on current inputs, accurately modeling temporal changes in human poses.

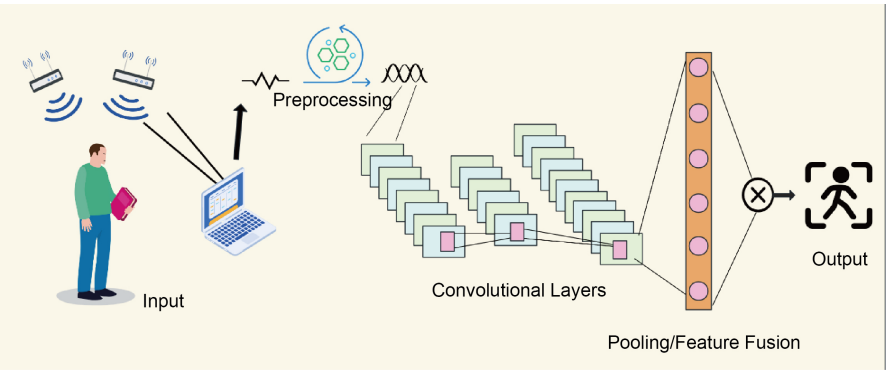

Figure 1. Preliminary human pose representation by CNN.

3.4. Model Training and Optimization

3.4.1. Multimodal Data-Driven Model Training Process

Using preprocessed WiFi signals and fused multimodal data as input, a deep learning model is constructed based on selected algorithms. To enable the model to accurately learn human pose and position patterns, researchers collected a large amount of diverse labeled data covering different indoor scenarios (living rooms, offices) [30], various human actions (walking, jumping, sitting, standing), and different lighting conditions, personnel numbers, and interactions.

During training, CNN extracts spatial features (e.g. limb positions, movement amplitudes) via layer-wise convolution with sliding kernels, providing a basis for subsequent analysis. Transformer processes CNN-extracted features using self-attention to strengthen inter-feature connections. LSTM takes features processed by CNN and Transformer as input, using gating mechanisms to handle sequential features and learn temporal variation patterns of human poses, enabling recognition of complex action sequences.

Through repeated training on large labeled datasets, the model continuously optimizes parameters, gradually learning human pose and position patterns, and improving performance in multi-person real-time 3D pose recognition tasks, laying a foundation for intelligent and accurate 3D human pose recognition.

3.4.2. Model Optimization Strategies for Complex Scenarios

During model training, optimization is performed to enhance positioning and pose recognition performance in complex scenarios.

Network parameter optimization is critical. For CNN, convolution kernel size, quantity, and pooling methods matter: small kernels capture details, and large kernels extract global features. For pooling, max pooling retains salient features, while average pooling smooths extraction; alternating use is recommended. For Transformer, adjusting parameters like the number of multi-head attention heads and hidden layer dimensions optimizes long-range dependency capture. For LSTM, adjusting layer numbers, neuron connections, and gate weights enhances learning capabilities.

Optimization algorithms and hyperparameters are also important. Adam converges fast and stably, while SGD is simple but learning rate-dependent; choices

should be based on data and training conditions. Learning rate decay strategies and regularization parameters (L1/L2) prevent overfitting.

Introducing attention mechanisms enhances model performance. In multi-person WiFi data processing, attention calculates feature weights to focus on key features and suppress interference. Adding attention modules to the model dynamically adjusts weights, enabling more accurate pose and position recognition and improving accuracy and efficiency. The attention mechanism principle is shown in Figure 2.

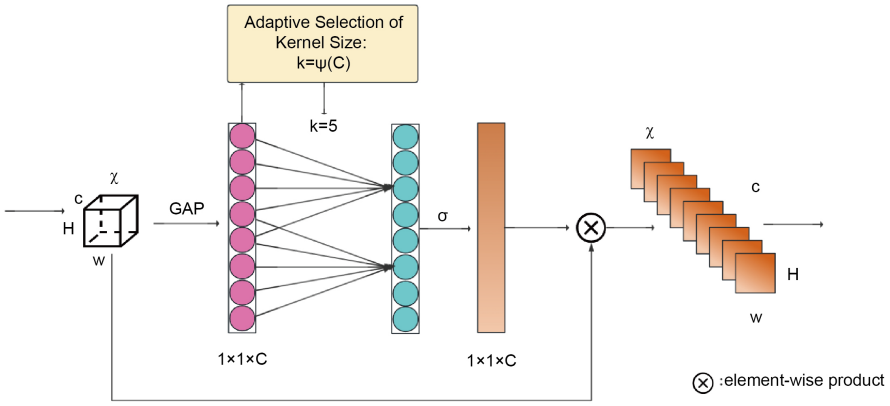

Figure 2. Schematic of the attention mechanism.

The section systematically describes the methodology for human behavior recognition, including data collection, preprocessing, algorithm design, and model training. By integrating signal processing techniques (e.g. phase unwrapping, Wiener filtering) and a hybrid deep learning architecture (CNN-Transformer-LSTM), the proposed framework addresses critical challenges in existing WiFi-based pose estimation, such as noise-sensitive phase signals and limited spatiotemporal feature learning.

However, despite these advancements, the outlined methods face inherent limitations. Existing approaches often rely on simplified assumptions about signal propagation or human motion, which may not fully generalize to highly dynamic environments with strong multipath interference. Additionally, while the hybrid model enhances feature representation, the computational complexity poses challenges for real-time deployment on resource-constrained devices.

Signal processing and model training are invariably integral to pose recognition. Given that the suitability of specific methods is largely determined by the application, selecting and adapting the processing approach to address specific requirements is crucial for the development of real-world systems.

4. Typical Applications

4.1. Gesture Recognition

Gesture recognition, a core technology in human-computer interaction, holds significant value in smart devices, virtual reality, and other fields. Traditional meth-

ods relying on specialized sensors or cameras suffer from high deployment costs and line-of-sight limitations. In recent years, contactless gesture recognition based on WiFi signals has become a research hotspot due to its lack of need for extra hardware and wide coverage. Below, two representative studies are discussed, focusing on position-agnostic sensing and fine-grained finger gesture recognition.

Gao *et al*. [34] proposed a position-agnostic sensing technology to address the issue of signal features in traditional WiFi gesture recognition being affected by user position and orientation. They innovatively shifted the observation perspective from "transceiver view" to "hand view" and introduced Motion Navigation Primitive (MNP) as a core feature. MNP extracts position-agnostic invariant features by analyzing patterns of hand movement direction changes in gestures. Experiments show the system achieves an average recognition accuracy of over 92% for 10 gestures across different positions, orientations, and environments, significantly outperforming traditional methods and providing key support for universal interaction in complex scenarios.

Tan *et al*. [35] designed the WiFinger system, focusing on single-finger micro-action recognition using CSI from commercial WiFi devices to achieve fine-grained gesture sensing. To address environmental noise and individual differences, the system employs multipath mitigation and wavelet denoising to filter dynamic environmental interference and retain subtle signal changes caused by gestures. It uses Principal Component Identification to extract inherent gesture features and selects subcarriers sensitive to finger movements to handle inter-user differences effectively. In home and office environments, WiFinger achieves an average recognition accuracy of over 93% for 8 finger gestures (e.g. zoom, flip, slide) and maintains high robustness in Non-Line-of-Sight (NLOS) scenarios. This study breaks through traditional methods' reliance on high sampling rates and specialized hardware, proving the feasibility of commercial WiFi devices in fine-grained gesture recognition.

In summary, WiFi-based gesture recognition has made significant progress in position agnosticism and fine-grained recognition through signal processing and feature extraction innovations, laying a foundation for practical deployment of interaction systems. Future research could further integrate multimodal fusion and deep learning to improve recognition accuracy and generalization in complex scenarios.

4.2. Vital Sign Monitoring

Monitoring vital signs during sleep is crucial for health assessment and disease diagnosis. Contactless sensing based on WiFi signals provides low-cost, non-invasive solutions for monitoring key indicators like respiration and heart rate by mining fine-grained CSI features. Below, two key studies are systematically discussed, from single-user fine monitoring to multi-user collaborative sensing, outlining technological advancements and application breakthroughs in this field, such as [36].

Liu *et al*. [36] proposed the first sleep monitoring system based on commercial WiFi devices, using CSI sensitivity to human micro-movements to synchronously monitor respiratory rate, heart rate, and sleep posture. The system uses Hampel filters and moving average filtering to remove noise, selects sensitive channels via subcarrier variance, and extracts sinusoidal features of respiratory cycles and high-frequency components of heart rates via time-frequency analysis. Experiments show that in single-user scenarios, respiratory rate monitoring error is less than 0.5 breaths/minute, heart rate error is within ±3 beats/minute, and four sleep postures (supine, side-lying, etc.) are recognized with over 90% accuracy. Its core advantage lies in requiring no extra hardware, relying solely on existing WiFi devices, making it feasible for long-term home health management, especially for elderly users with low wearable device compliance. This study demonstrates the effectiveness of commercial WiFi devices in contactless heart rate monitoring, providing technical support for early screening of diseases like sleep apnea and heart rate abnormalities.

For monitoring in multi-person shared spaces (e.g. wards, dormitories), Zeng *et al*. [37] developed the MultiSense system, achieving multi-user respiration signal separation from a single pair of WiFi devices via Blind Source Separation (BSS) and Independent Component Analysis (ICA) for the first time. The system uses the linear mixing characteristics of multi-antenna received signals to build signal models, remove background noise and phase offsets, and distinguish different individuals' respiratory rates via K-means clustering. Experimental results show 92.7% respiratory rate recognition accuracy in 3-person monitoring scenarios, with robustness to personnel position changes (e.g. within 50 cm movement). Compared to traditional spectral analysis methods, MultiSense breaks through the "blind spot" limitation, effectively monitoring even in weak signal reflection areas, and offering efficient solutions for group health monitoring in nursing homes, post-disaster rescue, and other scenarios. Technical comparisons and evolution are shown in Table 1.

Table 1. Technical comparison and evolution of vital sign monitoring.

	Single-user Monitoring [36]	Multi-user Sensing [37]
Core Method	Time-freq analysis, subcarrier screening, peak detection	Blind Source Separation (BSS), Independent Component Analysis (ICA)
Monitoring Target	Respiratory rate, heart rate, sleep posture	Multi-person resp sync monitoring, up to 4 people
Signal Processing	Band-pass filter to separate resp & heart rate signals	Multi-antenna linear mixing modeling, remove phase offset & background noise
Environmental Adaptability	Supports NLOS, 10m distance, multiple sleep postures	Resist furniture movement, AC vibration noise, strong positioning robustness
Accuracy	Respiratory rate error < 0.5 times/minute, heart rate ± 3 times/minute	3-person: 92.7% accuracy, avg absolute error 0.73/min
Hardware Dependence	Single-pair commercial WiFi (router + laptop)	Same, multi-antenna for better signal separation

4.3. Personnel Identification

In IoT and smart environments, personnel identification, as a core technology for security authentication and personalized services, faces challenges such as user compliance, privacy protection, and environmental adaptability. Traditional vision-based or biometric-based methods (e.g. facial recognition, fingerprint scanning) require active user cooperation and pose privacy risks, while wireless gait-based recognition, though passive, requires users to walk several meters along fixed paths, leading to high time costs and limited application scenarios. Addressing these issues, Wang *et al*.'s [38] WiPIN system was a new paradigm of passive, operation-free personnel identification. Its core lies in using unique signal distortions caused by individual physiological characteristics (e.g. body shape, body fat rate, muscle distribution) when WiFi signals penetrate the human body for identity differentiation. Users only need to stand still for ~200 ms for identification, significantly improving user-friendliness and system practicality.

Experimental validation shows WiPIN achieves 92% recognition accuracy in 30-person tests, maintaining stable performance as user scale increases (2 - 30 people). Its robustness is demonstrated by maintaining >90% accuracy over 15 days via periodic training data updates, achieving 94% intra-category and minimum of 77% cross-category recognition accuracy. This study not only expands WiFi signals' application boundaries in passive perception, but also proves the feasibility and superiority of wireless signal-human feature integration for recognition technology in real-world deployment.

4.4. Summary

A summary of the above application scenarios is shown in Table 2. WiFi-based contactless sensing technology has demonstrated groundbreaking application value across multiple fields by innovating signal processing and feature extraction, and building low-cost, high-robustness intelligent sensing systems. This technology system, centered on "contactless sensing-multi-scenario adaptation-efficient

Table 2. Summary of typical application scenarios.

	Core Objective	Key Technology	Performance Index	Typical Scenarios
Gesture Recognition	Recognize digital gestures (0 - 9) & finger motions (zoom, slide)	MNP, multipath signal proc., sub-carrier screening	Digital gesture acc. >92%, finger motion acc. >93%	Smart device int., VR, smart home
Vital Sign Monitoring	Single-user: sleep resp/HR/posture; multi-user: resp sync mon.	Time-freq analysis, BSS, ICA	Single-user: resp err. <0.5/min, HR ±3/min, posture acc. >90%; multi-user (3): resp acc. 92.7%	Home health mgt, nursing homes, hospital wards
Passive Personnel Identification	No-op identity auth (physiological feature signal distortion)	Signal filtering, feature extraction, SVM classification	30-person acc. 92%, single-time recog. <300 ms	Security auth, smart space services

recognition", lays a technical foundation for smart interaction, medical monitoring, intelligent security, and other fields. Future research could integrate multimodal fusion and deep learning to further enhance generalization in complex environments and promote large-scale application.

5. Existing Challenges

5.1. Complex Environment Effect

In complex indoor environments, WiFi signals are susceptible to multipath effects, electromagnetic interference, and other factors, causing increased CSI data fluctuations and noise. Despite using various filtering and phase correction algorithms, signal processing remains suboptimal in scenarios with strong interference sources or signal blockages. Achieving precise phase correction and deep noise removal remains challenging, affecting human feature extraction accuracy and limiting 3D pose recognition precision.

5.2. Challenges in Multimodal Data Fusion

Fusing WiFi signal features with camera video features is complicated by differences in data characteristics, dimensions, and sampling frequencies, making correlation and integration difficult. Millisecond-level errors in data synchronization and inconsistent feature representations can lead to redundant fused data or loss of critical information, hindering comprehensive and accurate reflection of human states and reducing model learning efficiency and recognition performance.

5.3. Optimization Bottlenecks in Model Training

During model training, hyperparameter tuning for deep learning architectures is complex and time-consuming. Adjusting parameters like CNN convolution kernels, LSTM network structures, and attention mechanism scopes requires extensive experimental exploration. Optimization algorithms like Adam and SGD, affected by data distribution and gradient vanishing, may get trapped in local optima for specific datasets/tasks, slowing model convergence, weakening generalization, and impeding accuracy improvements for multi-person complex scenarios.

5.4. Limitations of WiFi-Based Sensing

WiFi sensing, despite privacy and cost benefits, faces limitations. First, signal reliability is challenged by multipath interference in complex environments, impacting feature extraction accuracy, especially for fine details and rapid movements. Second, limited spatial/temporal resolution due to lower frequencies hinders capturing high-fidelity human contours and subtle motion. Signal superposition in multi-person scenarios further complicates individual separation. Third, environmental adaptability is bounded by penetration limits and signal attenuation; changes necessitate recalibration. Finally, compared to vision, WiFi lacks high-fidelity visual features, and precision compared to specialized hardware.

Multimodal fusion is needed, leveraging WiFi's privacy benefits while integrating visual/radar data to compensate for limitations and balance practicality and performance.

6. Future Development Suggestions

6.1. Innovations in Signal Processing Technology

Future developments should focus on signal processing technologies.

Develop new adaptive algorithms: Deep learning-driven signal processing models intelligently capture signal patterns based on environmental changes, strengthening filtering/denoising during multipath interference and precise phase offset correction to stabilize human feature extraction. These algorithms enhance anti-interference capabilities, ensuring feature extraction accuracy, and improving overall human action recognition performance in complex environments.

6.2. Efficient Multimodal Application Strategies

Develop intelligent fusion algorithms to automatically learn feature mapping and weight allocation via deep learning, enabling seamless fusion of WiFi and video data. Utilize Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) to mine latent data correlations, align spatiotemporal features, unify feature representation formats, improve fusion efficiency and data quality, provide accurate human state information for model input, and enhance recognition performance.

6.3. New Approaches to Model Training Optimization

Introduce Neural Architecture Search (NAS) technology to automatically design optimal network architectures, allocating computing resources, determining layer connections, and configuring convolution kernels based on task requirements. Combine adaptive learning rate adjustment strategies, gradient clipping, and novel regularization methods to mitigate gradient issues and accelerate convergence to global optima. Apply few-shot learning and transfer learning to leverage pre-trained model knowledge, reduce data dependency, and rapidly optimize models for improved action recognition accuracy and efficiency in complex scenarios.

7. Conclusion

This paper presents a real-time multi-person pose recognition system based on WiFi signals and deep learning neural networks using ubiquitous WiFi devices. The system accurately recognizes human pose in complex environments, enabling precise analysis of human movements. During experiments, large volumes of WiFi signal data were collected, preprocessed, and feature-extracted with deep learning neural networks. Extensive experimental validation and testing show that the constructed deep learning neural network performs well in recognizing human pose, effectively improving the accuracy and efficiency of human action recognition in complex environments. This technology creates more convenient and intelligent

experiences for fields, such as intelligent environmental perception, virtual reality, and human-computer interaction.

Acknowledgements

The work is funded by the foundation of the Innovation and Entrepreneurship Training Program for College Students (202410424048).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., *et al.* (2023) Deep Learning-Based Human Pose Estimation: A Survey. *ACM Computing Surveys*, **56**, 1-37. https://doi.org/10.1145/3603618

[2] Miao, F., Huang, Y., Lu, Z., Ohtsuki, T., Gui, G. and Sari, H. (2025) Wi-Fi Sensing Techniques for Human Activity Recognition: Brief Survey, Potential Challenges, and Research Directions. *ACM Computing Surveys*, **57**, 1-30. https://doi.org/10.1145/3705893

[3] Liu, W., Bao, Q., Sun, Y. and Mei, T. (2022) Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *ACM Computing Surveys*, **55**, 1-41. https://doi.org/10.1145/3524497

[4] Yousefi, S., Narui, H., Dayal, S., Ermon, S. and Valaee, S. (2017) A Survey on Behavior Recognition Using WiFi Channel State Information. *IEEE Communications Magazine*, **55**, 98-104. https://doi.org/10.1109/mcom.2017.1700082

[5] Chen, C., Zhou, G. and Lin, Y. (2023) Cross-Domain WiFi Sensing with Channel State Information: A Survey. *ACM Computing Surveys*, **55**, 1-37. https://doi.org/10.1145/3570325

[6] Ma, Y., Zhou, G. and Wang, S. (2019) WiFi Sensing with Channel State Information. *ACM Computing Surveys*, **52**, 1-36. https://doi.org/10.1145/3310194

[7] Zhou, Y., Xu, C., Zhao, L., Zhu, A., Hu, F. and Li, Y. (2022) CSI-Former: Pay More Attention to Pose Estimation with WiFi. *Entropy*, **25**, Article 20. https://doi.org/10.3390/e25010020

[8] Zhou, F., Zhu, G., Li, X., Li, H. and Shi, Q. (2023) Towards Pervasive Sensing: A Multimodal Approach via CSI and RGB Image Modalities Fusion. *Proceedings of the* 3*rd ACM MobiCom Workshop on Integrated Sensing and Communications Systems*, Madrid, 6 October 2023, 25-30. https://doi.org/10.1145/3615984.3616505

[9] Lai, Y., Horng, T., Su, W. and Lin, J. (2024) Wi-Fi-Based Posture Imaging Radar for Vital Sign Monitoring and Fall Detection. *IEEE Transactions on Microwave Theory and Techniques*, **72**, 6062-6071. https://doi.org/10.1109/tmtt.2024.3381626

[10] Lin, K., Wang, L. and Liu, Z. (2021) End-to-End Human Pose and Mesh Reconstruction with Transformers. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 1954-1963. https://doi.org/10.1109/cvpr46437.2021.00199

[11] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L. and Lu, C. (2021) HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 3382-3392. https://doi.org/10.1109/cvpr46437.2021.00339

[12] Jin, Z.Q. (2023) Research on 3D Human Body Reconstruction Based on Single Frame Image and SMPL. Master's Thesis, Zhejiang Gongshang University.

[13] Adib, F., Hsu, C., Mao, H., Katabi, D. and Durand, F. (2015) Capturing the Human Figure through a Wall. *ACM Transactions on Graphics*, **34**, 1-13. https://doi.org/10.1145/2816795.2818072

[14] Zhao, M., Li, T., Alsheikh, M.A., Tian, Y., Zhao, H., Torralba, A., *et al*. (2018) Through-Wall Human Pose Estimation Using Radio Signals. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7356-7365. https://doi.org/10.1109/cvpr.2018.00768

[15] Zhao, M., Liu, Y., Raghu, A., Zhao, H., Li, T., Torralba, A., *et al*. (2019) Through-Wall Human Mesh Recovery Using Radio Signals. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 10112-10121. https://doi.org/10.1109/iccv.2019.01021

[16] Li, G., Zhang, Z., Yang, H., Pan, J., Chen, D. and Zhang, J. (2020) Capturing Human Pose Using Mmwave Radar. 2020 *IEEE International Conference on Pervasive Computing and Communications Workshops* (*PerCom Workshops*), Austin, 23-27 March 2020, 1-6. https://doi.org/10.1109/percomworkshops48775.2020.9156151

[17] Kwon, S.M., Yang, S., Liu, J., Yang, X., Saleh, W., Patel, S., *et al*. (2019) Demo: Hands-Free Human Activity Recognition Using Millimeter-Wave Sensors. 2019 *IEEE International Symposium on Dynamic Spectrum Access Networks* (*DySPAN*), Newark, 11-14 November 2019, 1-2. https://doi.org/10.1109/dyspan.2019.8935665

[18] Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., *et al*. (2021) mmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave. *Proceedings of the* 19*th Annual International Conference on Mobile Systems, Applications, and Services*, 24 June-2 July 2021, 269-282. https://doi.org/10.1145/3458864.3467679

[19] Gu, T., Fang, Z., Yang, Z., Hu, P. and Mohapatra, P. (2019) mmSense: Multi-Person Detection and Identification via Mmwave Sensing. *Proceedings of the* 3*rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, Los Cabos, 25 October 2019, 45-50. https://doi.org/10.1145/3349624.3356765

[20] Wang, Y., Guo, L., Lu, Z., Wen, X., Zhou, S. and Meng, W. (2021) From Point to Space: 3D Moving Human Pose Estimation Using Commodity WiFi. *IEEE Communications Letters*, **25**, 2235-2239. https://doi.org/10.1109/lcomm.2021.3073271

[21] Halperin, D., Hu, W., Sheth, A. and Wetherall, D. (2011) Tool Release: Gathering 802.11n Traces with Channel State Information. *ACM SIGCOMM Computer Communication Review*, **41**, 53-53. https://doi.org/10.1145/1925861.1925870

[22] Hernandez, S.M. and Bulut, E. (2020) Lightweight and Standalone IoT Based WiFi Sensing for Active Repositioning and Mobility. 2020 *IEEE* 21*st International Symposium on "A World of Wireless, Mobile and Multimedia Networks"* (*WoWMoM*), Cork, 31 August-3 September 2020, 277-286. https://doi.org/10.1109/wowmom49955.2020.00056

[23] Yang, J., Zou, H., Jiang, H. and Xie, L. (2018) Device-Free Occupant Activity Sensing Using WiFi-Enabled IoT Devices for Smart Homes. *IEEE Internet of Things Journal*, **5**, 3991-4002. https://doi.org/10.1109/jiot.2018.2849655

[24] Wang, F., Zhou, S., Panev, S., Han, J. and Huang, D. (2019) Person-In-WiFi: Fine-Grained Person Perception Using WiFi. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5451-5460. https://doi.org/10.1109/iccv.2019.00555

[25] Wang, X., Gao, L. and Mao, S. (2015) PhaseFi: Phase Fingerprinting for Indoor Lo-

calization with a Deep Learning Approach. 2015 *IEEE Global Communications Conference* (*GLOBECOM*), San Diego, 6-10 December 2015, 1-6. https://doi.org/10.1109/glocom.2015.7417517

[26] Jiang, W., Xue, H., Miao, C., Wang, S., Lin, S., Tian, C., *et al*. (2020) Towards 3D Human Pose Construction Using WiFi. *Proceedings of the* 26*th Annual International Conference on Mobile Computing and Networking*, London, 21-25 September 2020, 1-4. https://doi.org/10.1145/3372224.3380900

[27] Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y. and Yang, J. (2021) 3D Human Pose Estimation Using WiFi Signals. *Proceedings of the* 19*th ACM Conference on Embedded Networked Sensor Systems*, Coimbra, 15-17 November 2021, 363-364. https://doi.org/10.1145/3485730.3492871

[28] Han, B., Wang, L., Lu, X., Meng, J. and Zhou, Z. (2023) Cross-Modal Meta-Learning for WiFi-Based Human Activity Recognition. *Proceedings of the* 29*th Annual International Conference on Mobile Computing and Networking*, Madrid, 2-6 October 2023, 1-3. https://doi.org/10.1145/3570361.3615754

[29] Wang, Y., Ren, Y. and Yang, J. (2024) Multi-Subject 3D Human Mesh Construction Using Commodity WiFi. *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, **8**, 1-25. https://doi.org/10.1145/3643504

[30] Wang, Y., Ren, Y., Chen, Y. and Yang, J. (2022) Wi-Mesh: A WiFi Vision-Based Approach for 3D Human Mesh Construction. *Proceedings of the* 20*th ACM Conference on Embedded Networked Sensor Systems*, Boston, 6-9 November 2022, 362-376. https://doi.org/10.1145/3560905.3568536

[31] Wang, Y., Ren, Y., Chen, Y. and Yang, J. (2022) A WiFi Vision-Based 3D Human Mesh Reconstruction. *Proceedings of the* 28*th Annual International Conference on Mobile Computing and Networking*, Sydney, 17-21 October 2022, 814-816. https://doi.org/10.1145/3495243.3558247

[32] Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y. and Yang, J. (2022) GoPose: 3D Human Pose Estimation Using WIFI. *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, **6**, 1-25. https://doi.org/10.1145/3534605

[33] Ren, Y., Wang, Z., Tan, S., Chen, Y. and Yang, J. (2021) Winect: 3D Human Pose Tracking for Free-form Activity Using Commodity WIFI. *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, **5**, 1-29. https://doi.org/10.1145/3494973

[34] Gao, R., Zhang, M., Zhang, J., Li, Y., Yi, E., Wu, D., *et al*. (2021) Towards Position-Independent Sensing for Gesture Recognition with Wi-Fi. *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, **5**, 1-28. https://doi.org/10.1145/3463504

[35] Tan, S. and Yang, J. (2016) WiFinger: Leveraging Commodity WiFi for Fine-Grained Finger Gesture Recognition. *Proceedings of the* 17*th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Paderborn, 5-8 July 2016, 201-210. https://doi.org/10.1145/2942358.2942393

[36] Liu, J., Chen, Y., Wang, Y., Chen, X., Cheng, J. and Yang, J. (2018) Monitoring Vital Signs and Postures during Sleep Using WiFi Signals. *IEEE Internet of Things Journal*, **5**, 2071-2084. https://doi.org/10.1109/jiot.2018.2822818

[37] Zeng, Y., Wu, D., Xiong, J., Liu, J., Liu, Z. and Zhang, D. (2020) MultiSense: Enabling Multi-person Respiration Sensing with Commodity WiFi. *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, **4**, 1-29. https://doi.org/10.1145/3411816

[38] Wang, F., Han, J., Lin, F. and Ren, K. (2019) WiPIN: Operation-Free Passive Person Identification Using Wi-Fi Signals. 2019 *IEEE Global Communications Conference* (*GLOBECOM*), Waikoloa, 9-13 December 2019, 1-6. https://doi.org/10.1109/globecom38437.2019.9014226