WiFi-Based Gesture Perception: Enabling Non-Contact Hand Gesture Recognition for AI System

Weiyi Jin*
weiyi.jin@temple.edu
Temple University
Philadelphia, Pennsylvania, USA

Abstract

This work explores a solution of enabling machines to perceive and interpret human gestures through ambient electromagnetic signals. We develop a perceptual system that leverages Wi-Fi Channel State Information (CSI) as a sensing modality for non-contact hand gesture recognition. The system implements a multi-stage perception pipeline—from signal acquisition and sensory preprocessing to pattern recognition and classification—achieving an F1-score of 0.97 across six gesture classes. This work demonstrates how AI systems can extend perceptual capabilities beyond traditional visual and tactile sensing, enabling machines to "see" human motion through radio frequency variations. We analyze the system from an AI perception perspective, examining feature extraction, pattern discrimination, and the learned representations that enable robust gesture recognition.

Keywords

N/A

ACM Reference Format:
Weiyi Jin. 2025. WiFi-Based Gesture Perception: Enabling Non-Contact Hand Gesture Recognition for AI System. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Perception in Artificial Intelligence

Perception represents a foundational capability in artificial intelligence—the ability to acquire, process, and interpret sensory information from the environment to form meaningful representations that enable intelligent behavior. While biological systems utilize vision, hearing, touch, and other modalities to perceive their surroundings, AI systems must design and implement artificial perceptual mechanisms that can extract relevant information from diverse signal sources.

Traditional AI perception systems have primarily focused on visual (camera-based) and auditory sensing. However, the space of possible sensing modalities is much richer. This work explores *radio frequency (RF) perception*—specifically, using Wi-Fi Channel State Information (CSI) as a sensory modality for perceiving human gestures. This represents a novel form of machine perception that extends AI's sensory capabilities beyond conventional approaches.

1.2 The Gesture Perception Problem

Gesture recognition exemplifies a classical perception problem in AI: given continuous sensory input (in this case, WiFi signal variations), the system must:

- **Sense**: Acquire raw sensory data from the environment
- **Process**: Transform raw signals into meaningful features
- **Recognize**: Identify patterns corresponding to discrete gesture categories
- **Classify**: Map observations to semantic labels (gesture types)

This perception pipeline parallels how biological systems process sensory information. Just as human visual perception transforms retinal patterns into recognized objects, our system transforms RF signal patterns into recognized gestures.

1.3 Why WiFi-Based Perception

Wi-Fi Channel State Information (CSI) offers unique advantages as a perceptual modality:

- **Ubiquitous Sensing**: Leverages existing WiFi infrastructure, requiring no additional sensors
- **Privacy-Preserving**: Unlike cameras, RF sensing does not capture visual appearance
- **Non-Invasive**: Requires no worn devices or user instrumentation
- **Robust to Environmental Conditions**: Functions in darkness, occlusion, and varied lighting
- **Penetrating Capability**: RF signals can penetrate obstacles, enabling through-wall sensing

As illustrated in Figure 1, WiFi transmitters (TX) and receivers (RX) continuously exchange signals. When a user performs hand gestures, these movements create variations in signal propagation that manifest as changes in the CSI data. The challenge is to build an AI perception system that can learn to interpret these signal variations as meaningful gesture patterns.

1.4 Research Contributions

This work makes the following contributions to AI perception:

- **Perceptual System Design**: We design and implement a complete perception pipeline for RF-based gesture recognition, from raw signal acquisition to high-level gesture classification

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

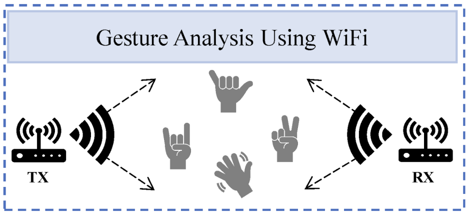


Figure 1: Example Usage of Wi-Fi CSI for Hand Gesture Recognition

- **Feature Learning**: We develop preprocessing methods that enhance perceptually-relevant signal variations while suppressing noise, analogous to early sensory processing in biological systems
- **Pattern Recognition**: We apply deep residual learning to automatically discover discriminative features for gesture patterns, achieving 97% F1-score
- **Perceptual Analysis**: We analyze what patterns the system learns, where perception fails (confusion between similar gestures), and how preprocessing affects perceptual capabilities

2 Related Work

2.1 AI Perception Systems

Perception in AI has evolved from hand-crafted feature extractors to learned representations. Classical approaches relied on domain experts to design features (e.g., SIFT for vision, MFCCs for audio), while modern deep learning enables end-to-end perceptual learning where features emerge automatically from data [7].

2.2 WiFi-Based Sensing as Perception

Wi-Fi CSI has emerged as a powerful perceptual modality for human activity recognition [2, 4]. These systems implement various stages of the perception pipeline:

- **Signal Processing Layer**: Filtering and denoising to enhance signal-to-noise ratio [5]
- **Feature Extraction**: Converting raw CSI into representations suitable for pattern recognition [3]
- **Pattern Recognition**: Applying machine learning to recognize activities from features [1]
- **Domain Adaptation**: Enabling perception to generalize across environments [6]

Wang et al. [1] developed Widar3.0, demonstrating that velocity profiles extracted from CSI can serve as domain-independent perceptual features. This mirrors biological perception, where motion patterns (optic flow) provide robust cues regardless of specific visual appearance.

2.3 Deep Learning for Perceptual Learning

Recent work has shown that deep neural networks can learn hierarchical perceptual representations from CSI data. Xu et al. [7] explored self-supervised learning techniques that enable perceptual systems to learn from unlabeled sensory data, reducing dependence on annotated training examples. This parallels how biological systems learn perceptual capabilities through unsupervised exposure to sensory stimuli.

3 Perception System Architecture

3.1 Perceptual Pipeline Overview

Our gesture perception system implements a multi-stage pipeline (Figure 2):

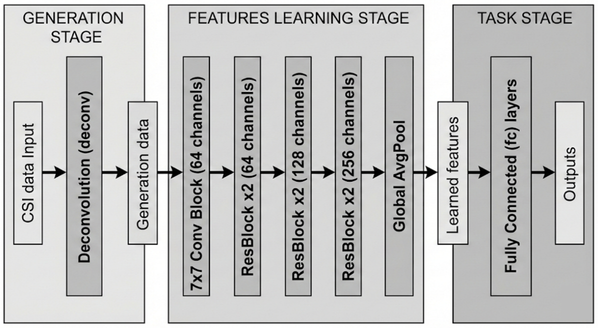


Figure 2: Perception System Architecture

(1) **Sensory Acquisition**: Raw CSI data capture
(2) **Preprocessing**: Signal filtering and feature selection
(3) **Feature Extraction**: Hierarchical representation learning via ResNet
(4) **Pattern Recognition**: Classification into gesture categories

This architecture mirrors the computational stages in biological perception, from early sensory processing to high-level object recognition.

3.2 Problem Formulation: Perception as Pattern Recognition

Formally, we define gesture perception as learning a mapping function:

$$f_\theta : \mathcal{X} \to \mathcal{Y}$$

where:

- $\mathcal{X} = \mathbb{R}^{T \times d}$ is the space of CSI observations (T timesteps, d features)
- $\mathcal{Y} = \{1, 2, ..., K\}$ is the discrete set of gesture categories
- θ represents learned perceptual parameters

The challenge is to learn θ such that f_θ maps similar sensory patterns to the same category while discriminating between different gesture types—a fundamental pattern recognition problem in AI.

3.3 Sensory Preprocessing: Enhancing Perceptual Signals

Raw sensory signals contain both information-bearing variations and noise. Effective perception requires isolating relevant signals, analogous to early sensory processing in biological systems (e.g., retinal preprocessing in vision).

We apply Butterworth filtering to extract low-frequency components that correspond to hand motion, removing high-frequency noise:

$$X_{filtered} = H(f) \cdot X_{raw}$$

where $H(f)$ is the Butterworth filter frequency response. This preprocessing enhances the signal-to-noise ratio, making subsequent pattern recognition more effective.

Additionally, we found that the real component of CSI carries more perceptually-relevant information than the imaginary component. This suggests that amplitude variations (captured by the real part) are more informative for gesture perception than phase variations.

3.4 Feature Learning: Hierarchical Perceptual Representations

We employ a Residual Network (ResNet) architecture to learn hierarchical perceptual features. ResNet's skip connections enable learning of increasingly abstract representations:

- **Early Layers**: Detect low-level signal patterns (edges, transitions)
- **Middle Layers**: Combine low-level features into motion primitives
- **Deep Layers**: Recognize complex gesture patterns

This hierarchical processing parallels the ventral visual stream in biological perception, where simple features combine to form increasingly complex representations.

The ResNet architecture consists of:

- Input convolutional layer (64 filters, stride 2)
- Three residual stages with progressive downsampling
- Global average pooling
- Fully connected classification layer

The residual connections enable gradient flow through deep networks, allowing the system to learn subtle perceptual distinctions.

4 Experimental Evaluation

4.1 Dataset and Perceptual Categories

We use the WIDAR 3.0 dataset [8], focusing on six distinct gesture categories (Figure 3):

- **Push & Pull**: Forward/backward motion
- **Sweep**: Lateral arm movement
- **Clap**: Repetitive hand collision
- **Slide**: Horizontal sliding motion
- **Draw-N**: Angular motion pattern (letter N)
- **Draw-O**: Circular motion pattern

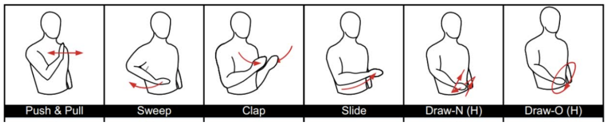


Figure 3: Dataset Gestures Overview

Each gesture produces a distinctive CSI signature. The dataset contains 1,500 samples per gesture (9,000 total), split 60% training, 20% validation, 20% testing.

4.2 Training the Perceptual System

We train the perception system using:

- **Optimizer**: Adam (learning rate: 0.001)
- **Loss Function**: Cross-entropy (measures perceptual confusion)
- **Epochs**: 30 (early convergence observed)
- **Batch Size**: 32

4.3 Perception Performance

Baseline System: Learning Without Preprocessing. The baseline ResNet without preprocessing achieved:

- Training accuracy: 89.41%
- Validation accuracy: 76.11%
- Test accuracy: 75.0%

The performance gap between training and testing suggests the perceptual system overfits to training data patterns, failing to generalize perceptual capabilities.

Enhanced Perception: The Role of Preprocessing. After applying Butterworth filtering and selecting the real CSI component, perception performance improved dramatically (Table 1):

Table 1: Perceptual Performance by Gesture Class

Gesture	Precision	Recall	F1-Score	Samples
Push & Pull	0.94	0.93	0.94	157
Sweep	0.93	0.94	0.94	154
Clap	1.00	1.00	1.00	135
Slide	0.98	1.00	0.99	160
Draw-N	1.00	0.96	0.98	141
Draw-O	0.98	0.99	0.99	153
Overall	**0.97**	**0.97**	**0.97**	**900**

This demonstrates that preprocessing—analogous to early sensory processing in biological systems—is crucial for effective perception.

Figure 4 illustrates the impact of filtering on perceptual signal quality:

The filtered signals exhibit clearer patterns, enabling more effective perceptual learning.

5 Perceptual Analysis

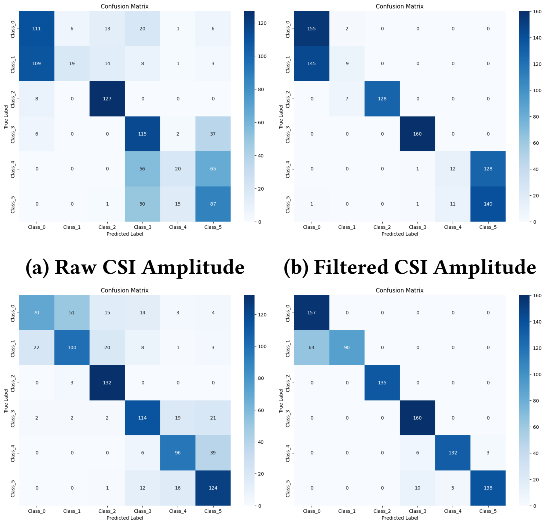
5.1 Where Perception Succeeds

The confusion matrix (Figure 5) reveals strong diagonal values, indicating robust perceptual discrimination. Most gestures achieve near-perfect recognition:

- **Clap**: 100% accuracy (distinctive repetitive pattern)
- **Slide**: 99% F1-score (clear horizontal motion signature)
- **Draw-O**: 99% F1-score (distinctive circular pattern)

5.2 Where Perception Fails: Perceptual Confusion

The system exhibits perceptual confusion between gestures with similar motion characteristics:



(a) Raw CSI Amplitude **(b) Filtered CSI Amplitude**

(c) Raw CSI Real Part **(d) Filtered CSI Real Part**

Figure 4: Impact of Preprocessing on Perceptual Signal Quality

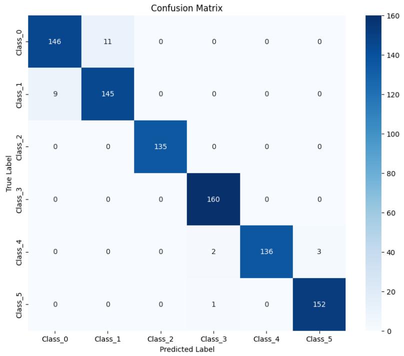


Figure 5: Confusion Matrix: Analyzing Perceptual Errors

Push & Pull vs. Sweep Confusion. The system misclassifies Push & Pull as Sweep in 11 cases. Both gestures involve horizontal arm motion:

- **Push & Pull**: Forward/backward motion along sagittal axis
- **Sweep**: Lateral motion along transverse axis

Both produce similar RF signatures with predominant motion in the horizontal plane. The perceptual system struggles to discriminate the axis of motion—a limitation analogous to motion ambiguities in human vision (e.g., the aperture problem).

Draw-N vs. Draw-O Confusion. Angular (Draw-N) and circular (Draw-O) motions occasionally confuse the system. This suggests:

- Smooth transitions in Draw-N resemble portions of Draw-O's circular motion
- The learned perceptual features may not fully capture angular vs. curved motion characteristics
- Additional temporal or geometric features may be needed for robust discrimination

5.3 Learning Dynamics: Training the Perceptual System

Figure 6 shows the learning curves:

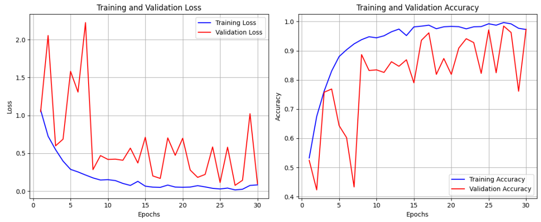


Figure 6: Perceptual Learning Dynamics

- **Training Loss**: Steady decrease indicates effective gradient-based learning
- **Validation Fluctuations**: Early instability stabilizes after epoch 10, suggesting the perceptual system initially learns noisy patterns before converging on robust features
- **Fast Convergence**: The system develops perceptual capabilities quickly, reaching peak performance by epoch 20

These dynamics resemble perceptual learning in biological systems, where initial learning is unstable but eventually converges to stable perceptual representations.

6 Discussion

6.1 The Importance of Sensory Preprocessing

Our results demonstrate that perceptual performance critically depends on preprocessing. Raw signals contain irrelevant variations that confuse pattern recognition. Effective perception requires filtering to isolate information-bearing components—a principle observed across biological sensory systems (retinal preprocessing, cochlear filtering, etc.).

This highlights a general principle in AI perception: **perceptual systems must actively process sensory input to extract relevant information, not merely record raw signals**.

6.2 Feature Modality Selection

The superior performance of real vs. imaginary CSI components reveals that not all signal aspects contribute equally to perception. The real component captures amplitude variations directly related to physical motion, while phase information (imaginary component) may be more susceptible to environmental noise.

This suggests: **effective perception requires understanding which signal modalities contain task-relevant information**, similar to how vision focuses on certain wavelengths or hearing emphasizes certain frequency ranges.

6.3 Perceptual Generalization vs. Overfitting

The baseline system's overfitting demonstrates a fundamental challenge in perceptual learning: memorizing training examples vs.

learning generalizable perceptual features. Preprocessing improved generalization by reducing dimensionality and removing spurious patterns.

This relates to the classical AI problem of **distinguishing relevant from irrelevant variations in sensory input**—the system must learn to ignore noise while capturing signal.

6.4 Hierarchical Perceptual Representations

ResNet’s layered architecture learns increasingly abstract perceptual features, from low-level signal patterns to high-level gesture representations. This hierarchical organization parallels biological perception and represents a key principle in AI: **complex perception emerges from compositional combination of simpler features**.

6.5 Limitations and Future Directions

Dataset Diversity. The single-user dataset limits perceptual generalization. Human perception develops through diverse sensory experiences; similarly, AI perception requires exposure to varied conditions. Future work should:

- Include multiple users with varying gesture styles
- Test across different environments (rooms, furniture layouts)
- Evaluate robustness to environmental dynamics

Perceptual Ambiguity. Some gestures remain perceptually ambiguous (Push/Sweep confusion). Biological systems resolve ambiguity through:

- **Multi-modal Perception**: Combining multiple sensory modalities
- **Temporal Context**: Using motion history to disambiguate
- **Attention Mechanisms**: Focusing on discriminative features

Future perception systems could incorporate these mechanisms.

Explainable Perception. Understanding *what* the system perceives remains challenging. Visualization techniques (activation maps, feature attribution) could reveal learned perceptual features, enabling:

- Debugging perceptual failures
- Validating learned representations
- Improving system design through understanding

Alternative Architectures. We tested AlexNet (Figure 7) but found ResNet superior. Other architectures (Transformers, attention-based models) may better capture temporal perceptual patterns. This suggests: **perceptual architecture design should match the structure of perceptual information in sensory signals**.

7 Conclusion

This work demonstrates that AI systems can develop novel forms of perception beyond traditional sensing modalities. By treating WiFi CSI as a perceptual signal, we enable machines to "see" human motion through electromagnetic variations—a form of perception unavailable to biological systems.

Key insights for AI perception include:

- **Preprocessing is Perceptual Processing**: Effective perception requires active signal transformation, not passive recording
- **Feature Learning Enables Perception**: Deep learning can automatically discover perceptual features from sensory data
- **Perceptual Errors Reveal Limitations**: Confusion patterns indicate what the system does/doesn’t perceive
- **Perceptual Generalization Requires Diversity**: Robust perception needs exposure to varied sensory conditions

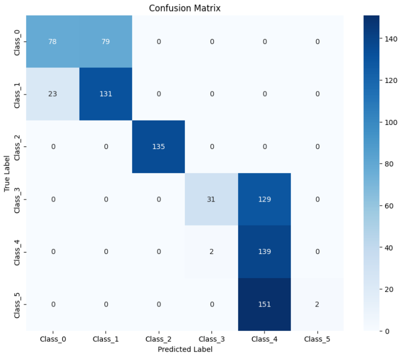


Figure 7: Alternative Architecture Performance: AlexNet

Our 97% F1-score demonstrates that RF-based gesture perception is feasible and effective. This opens possibilities for:

- Privacy-preserving human-computer interaction
- Ambient intelligence in smart environments
- Through-wall sensing for security and healthcare
- Multi-modal perception systems combining RF with vision/audio

As AI systems expand their perceptual capabilities, they can engage with the world through increasingly diverse sensory modalities, bringing us closer to truly intelligent machines that perceive and interpret their environment as richly as biological organisms.

Future Work: We aim to expand dataset diversity, explore attention mechanisms for resolving perceptual ambiguity, and develop multi-modal perception systems that integrate RF sensing with traditional modalities. Understanding and improving machine perception remains a fundamental challenge in achieving robust, generalizable artificial intelligence.

References

[1] Jiahui Chen, Haozhen Li, Lei Li, Boyuan Zhang, and Xinyu Gu. 2023. Robust WI-FI Enabled Human Activity Recognition Via Unsupervised Adversarial Domain Adaptation. In *2023 8th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*. IEEE, Beijing, China, 461–465. doi:10.1109/IC-NIDC59918.2023.10390678

[2] Neena Damodaran, Elis Haruni, Muyassar Kokhkharova, and Jörg Schäfer. 2020. Device Free Human Activity and Fall Recognition Using WiFi Channel State Information (CSI). *CCF Transactions on Pervasive Computing and Interaction* 2, 1 (March 2020), 1–17. doi:10.1007/s42486-020-00027-1

[3] Wei Guo, Shunsei Yamagishi, and Lei Jing. 2024. Human Activity Recognition via Wi-Fi and Inertial Sensors With Machine Learning. *IEEE Access* 12 (2024), 18821–18836. doi:10.1109/ACCESS.2024.3360490

[4] Guiping Lin, Weiwei Jiang, Sicong Xu, Xiaobo Zhou, Xing Guo, Yujun Zhu, and Xin He. 2023. Human Activity Recognition Using Smartphones With WiFi Signals. *IEEE Transactions on Human-Machine Systems* 53, 1 (Feb. 2023), 142–153. doi:10.1109/THMS.2022.3188726

[5] Julio C. H. Soto, Iandra Galdino, Egberto Caballero, Vinicius Ferreira, Débora Muchaluat-Saade, and Célio Albuquerque. 2022. A Survey on Vital Signs Monitoring Based on Wi-Fi CSI Data. *Computer Communications* 195 (Nov. 2022), 99–110. doi:10.1016/j.comcom.2022.08.004

[6] Xuyu Wang, Chao Yang, and Shiwen Mao. 2020. On CSI-Based Vital Sign Monitoring Using Commodity WiFi. *ACM Trans. Comput. Healthcare* 1, 3 (May 2020), 12:1–12:27. doi:10.1145/3377165

[7] Ke Xu, Jiangtao Wang, Hongyuan Zhu, and Dingchang Zheng. 2023. Self-Supervised Learning for WiFi CSI-Based Human Activity Recognition: A Systematic Study. arXiv:2308.02412 [eess] doi:10.48550/arXiv.2308.02412

[8] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 313–325.