

MULTIMODAL-BASED GAIT RECOGNITION USING COMMODITY WI-FI DEVICES

Sheng Chen, Fei Yang*, Aimin Pan, Zhewei Mei

Zhejiang Lab

ABSTRACT

Conventional approaches of human gait recognition are plagued by issues such as invasion on privacy, constraints in terms of space, limits in lighting conditions, and inconveniences associated with wearable gadgets. Herein, we explore the potential advantages of Wi-Fi signals and fuse two modalities, namely phase and spectrogram, of Wi-Fi Channel State Information (CSI) as gait features. This integration leads to the development of a robust human gait recognition system called MultiGaFi. Specifically, we extract time-frequency features of spectrograms through a well-designed network, while temporal features related to phase changes are captured through gated recurrent unit (GRU). Subsequently, the time-frequency features and temporal features are linearly fused to facilitate human gait recognition. We implement MultiGaFi on commodity Wi-Fi devices across four distinct indoor environments, and the empirical findings indicate that MultiGaFi achieves an impressive average accuracy of 98.11% within these specific indoor environments.

Index Terms— Wi-Fi Signal Processing, Human Gait Recognition, Channel State Information, Deep Learning

1. INTRODUCTION

Gait recognition, being a biometric identification technique, has gained significant interest in recent years and has been widely applied across various domains. The primary reliance of traditional gait recognition methods is on physical devices such as cameras [1,2], wearable devices [3], and pressure sensors [4]. Nevertheless, it is important to acknowledge that these approaches include inherent constraints, such as privacy invasion, reliance on wearable devices, and vulnerability to environmental factors. The limitations of traditional recognition methods have been addressed by Wi-Fi based recognition methods. Hence, a large number of human activity recognition systems [5–9] leveraging Wi-Fi signals have emerged, with gait recognition systems [10–13] gaining considerable prominence in this domain.

*Corresponding author. Email: yangf@zhejianglab.com

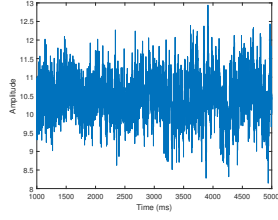
This work was supported by National Natural Science Foundation of China Grant (No. U22A6001) and Key Research Project of Zhejiang Lab (No. 2022PG0AC02).

In recent years, researchers have been exploring ways to enhance the performance of Wi-Fi based gait recognition systems. WifiU [12] is one of the most characteristic systems of previous works, which experientially extracts time-frequency gait features to train a machine learning model. However, WifiU exhibits several limitations as it restricts participants to traverse a singular predetermined trajectory and solely facilitates experimentation inside a singular environment. Besides, alternative approaches either prioritize on the temporal information of Wi-Fi Channel State Information (CSI) amplitude, neglecting the frequency-domain information [5, 10, 13], or exclusively examine the frequency-domain information of CSI [9, 11]. Consequently, these methods fail to comprehensively capture the multimodal features inherent in CSI data, leading to suboptimal performance. Given that human walking speed can influence the phase changes in CSI data [14], and that walking speed is a crucial component of human gait, we are inclined to include phase changes as a feature inside our system. Therefore, drawing inspiration from XModal-ID [15] and WiVi [16], which fuse cross-modal features, namely video and CSI, for recognition, our basic idea revolves upon the fusion of CSI phase features and CSI spectrogram features as inputs to enhance gait recognition.

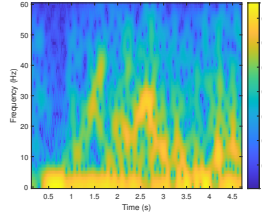
In this paper, we propose MulGaFi, a novel Multimodal-based Gait recognition system using Wi-Fi signals. MulGaFi employs short-time Fourier transform (STFT) to convert CSI data into spectrograms, facilitating extraction of features in the time-frequency domain. MulGaFi adopts phase recalibration to capture precise CSI phase changes, which enables the extraction of features in human walking speed. We design a gait recognition model which utilizes convolutional neural network (CNN) with SEBlock [17] and gated recurrent unit (GRU) to extract gait features in the time-frequency domain and temporal features related to phase changes separately. In addition, a fusion block is applied to fuse these features linearly, thereby improving gait recognition performance through the fusion of CSI spectrogram and CSI phase.

We highlight the main contributions as follows:

- (i) We propose a multimodal-based human gait recognition system based on Wi-Fi CSI, named MulGaFi, which utilizes two modalities of CSI to achieve higher performance gait recognition.
- (ii) We design a recognition network, which can extract



(a) Raw amplitude trace



(b) Log-Scaled spectrogram

Fig. 1. Raw CSI Amplitude vs. Spectrogram with STFT.

both spectrogram features in the time-frequency domain and temporal features related to phase changes, subsequently fusing them linearly for enhanced performance.

- (iii) We implement MulGaFi using commodity Wi-Fi devices and evaluate the performance through 8 subjects in four different environments. The experimental results indicate that MulGaFi can achieve an impressive average accuracy of 98.11% in these specific indoor environments.

2. SYSTEM DESIGN

The system is comprised of two modules, namely *CSI data pre-processing* and *gait recognition model*. The following sections describe the comprehensive processing procedures of each module.

2.1. CSI Data Pre-processing

2.1.1. Preliminaries

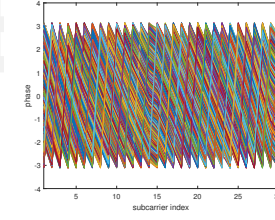
In the field of wireless communication, Wi-Fi CSI characterizes how wireless signals are transmitted from Wi-Fi transmitters to receivers through specific carrier frequencies. Specifically, CSI measurement can be represented as follow:

$$H(f, t) = \sum_{n=1}^N \alpha_n(f, t) e^{-j2\pi f \tau_n(t)}, \quad (1)$$

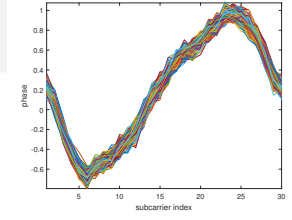
where f represents carrier frequency, N denotes the number of propagation paths, $\alpha_n(f, t)$ is the amplitude attenuation factor, which represents the amplitude attenuation on the n^{th} path, and $\tau_n(t)$ represents the delay on the n^{th} path. Based on this, the CSI phase can be obtained by adopting the inverse Fourier transform of the above CSI. Notably, both the amplitude and phase of CSI are subject to alterations induced by human activities.

2.1.2. CSI Amplitude Pre-processing

In the CSI amplitude pre-processing module, we interpolate, denoise, and reduce dimensionality on the raw CSI in turn



(a) Original phase w/o calibration



(b) Modified phase with calibration

Fig. 2. Comparison of CSI phase with/without calibration.

to get the modified CSI. Subsequently, STFT is applied to transform the modified CSI into spectrograms. The outlined processing procedures are as follows.

Interpolation: During the collection process of Wi-Fi signals, it is common to encounter packet losses caused by issues with transceiver equipment and uncertain propagation paths. Therefore, we utilize Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation [18], which aids in the restoration of temporal gaps between CSI packets, to ensure data integrity and reduce the adverse effects of packet losses.

Noise Filtering: Due to the fluctuations in transmission power and transmission rate of Wi-Fi signals [19], CSI measurements often contain a lot of noise, leading to random variation in the amplitude of CSI, as shown in Fig. 1(a). In order to address this issue, we first employ Hampel filter [20] to detect and remove outliers in the CSI amplitude. Next, we choose the wavelet denoising method to remove in-band noise in CSI amplitude. Furthermore, considering the frequency range of CSI fluctuations resulting from human activities typically lies between 20~60 Hz [12], we utilize Butterworth filter to eliminate high-frequency noise and preserve low-frequency signals associated with human walking.

Dimension Reduction: In our experiment, each CSI physical frame provides data for 90 subcarriers. Since the amplitude fluctuations among different subcarriers from the same Wi-Fi antenna pair tend to exhibit similar patterns, we utilize Principal Component Analysis (PCA) to reduce the dimensionality of these subcarriers, enhancing computational efficiency. Besides, based on the trade-off between PCA explained variance and our system's performance, we retain 12 principal components in our system.

Spectrogram Generation: Finally, we use STFT to convert the pre-processed 12-dimensional CSI data into 12-dimensional spectrograms, as illustrated in Fig. 1(b). These spectrograms could be regarded as 12-channel images, which serve as the input for our system.

2.1.3. CSI Phase Pre-processing

In the presence of Sampling Time Offset (STO) and Carrier Frequency Offset (CFO), the raw CSI phase is disordered. As

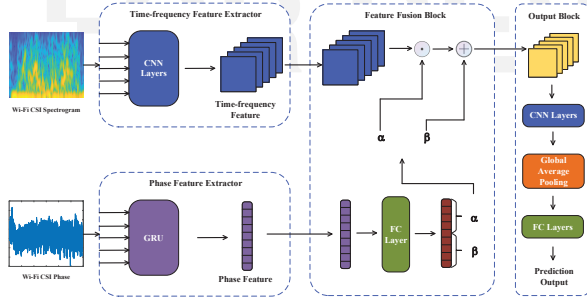


Fig. 3. Network architecture of Our System.

depicted in Fig. 2(a), which illustrates the phase transitions of 1000 successive CSI frames, it is evident that the CSI phase exhibits cyclic patterns and is folded. To address this issue, we first apply a phase unwrapping process to confirm the absence of phase discontinuities of π . Subsequently, to eliminate phase shifts caused by hardware limitations, we apply a linear transformation to the unwrapped CSI phase, mapping it to a different range. The result of this linear transformation on the phase is shown in Fig. 2(b), illustrating a heightened level of intuitiveness in the phase. Currently, the modified CSI phase serves as an additional input to our system.

2.2. Gait Recognition Model

As shown in Fig. 3, the model comprises of four network components: a time-frequency feature extractor, a phase feature extractor, a feature fusion block and an output block. The time-frequency feature extractor is tasked with extracting gait features from spectrograms within the time-frequency domain. The phase feature extractor is designed to receive CSI phase data and targets the extracting of temporal features that are relevant to human walking speed. The feature fusion block is used to linearly fuse the features acquired from the time-frequency and phase extractors. The output block performs gait recognition using the fused features.

2.2.1. Time-Frequency Feature Extractor

The time-frequency feature extractor, denoted as $\mathcal{F}(x_{spec}; \theta_{\mathcal{F}})$, takes 12-channel spectrogram data x_{spec} as its input. The architecture of the extractor is outlined in Table 1. It encompasses six CNN layers each with 3×3 kernel and 1×1 padding. Max pooling with 3×3 kernel and 2×2 stride is applied after every two layers. Besides, to improve feature representation and channel interdependencies, SEBlock is employed after every max pooling. Furthermore, each layer uses a LeakyReLU activator, and batch normalization is employed after the activator. The formula for time-frequency feature extractor is expressed as: $\hat{f} = \mathcal{F}(x_{spec}; \theta_{\mathcal{F}})$.

Module	Layer	Pooling
Time-frequency Feature Extractor CNN Layers	Conv (32@3×3)	Max Pooling
	Conv (64@3×3)	
	SEBlock	Max Pooling
	Conv (96@3×3)	
	Conv (96@3×3)	Max Pooling
	SEBlock	
	Conv (128@3×3)	Max Pooling
	Conv (128@3×3)	
Output Block CNN Layers	SEBlock	Global Average Pooling (64)
	Conv (128@3×3)	
	Conv (128@3×3)	
	Conv (64@3×3)	

Table 1. Layers Structure of Our Model.

2.2.2. Phase Feature Extractor

The phase feature extractor, denoted as $\mathcal{P}(x_{phase}; \theta_{\mathcal{P}})$, receives CSI phase x_{phase} as its input, where $x_{phase} \in \mathbb{R}^{L \times S}$. Here, L represents the number of subcarriers, and S denotes the length of the CSI phase data corresponding to a walking activity. This extractor employs two bi-directional GRU layers with 96 hidden units and a dropout rate set to 0.3. The result is the temporal feature capturing CSI phase changes. The formula for the phase feature extractor is expressed as: $\hat{p} = \mathcal{P}(x_{phase}; \theta_{\mathcal{P}})$.

2.2.3. Feature Fusion Block

To effectively fuse these two features of different modalities, we adopt a linear method that consumes fewer computational resources and time. Specifically, when dealing with the phase feature extracted by the phase feature extractor, as depicted in feature fusion block in Fig. 3, we employ a fully connected layer to map this feature to $(2 \times C)$ -dimensional representation (α and β), where C denotes the channel number of output \hat{f} from the time-frequency feature extractor.

Through the parameters α and β , we fuse these features in terms of element-wise multiplication and element-wise addition, which is expressed as: $\hat{b} = \hat{f} \odot \alpha + \beta$, where \hat{b} is output of feature fusion block, and \odot is element-wise multiplication. Hence, since α and β are independent to the time-frequency features, the feature fusion block can enhance our system's performance compared to using a single modal feature.

2.2.4. Output Block

Subsequently, the fused features are fed into an output block, denoted as $\mathcal{O}(\hat{b}; \theta_{\mathcal{O}})$. As depicted in Fig. 3 and detailed in Table 1, the block begins with three CNN layers each with 3×3 kernel. Each layer utilizes a LeakyReLU activator and applies batch normalization after the activator. Subsequently, we employ a fully connected layer with LeakyReLU activation before performing global average pooling. The resulting

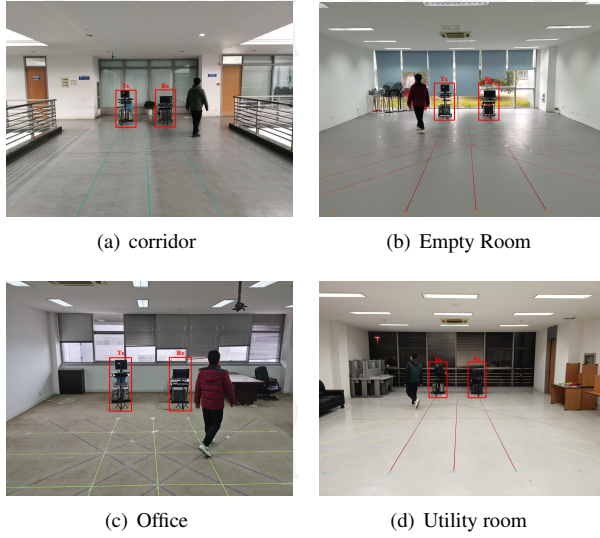


Fig. 4. Experimental scenarios.

output is considered the final prediction. The formula for output block is: $\hat{y} = \mathcal{O}(\hat{b}; \theta_{\mathcal{O}})$. Additionally, we use the cross-entropy loss function as our model's loss function.

3. EVALUATION

3.1. Experiment Setup

Experiment Scenarios: Considering the robustness of MulGaFi across various scenarios, we set up our system in four different scenarios, as illustrated in Fig. 4. These scenarios include a corridor, an empty room, an office, and a utility room. Within each scenario, we provide a pre-defined path grid for volunteers to walk, as shown in Fig. 5.

Data Collection: We collect CSI data using the CSI Tool released in [21] by Intel 5300 network card. Our setup involves a Wi-Fi transmitter equipped with an external omnidirectional antenna and a Wi-Fi receiver equipped with three external omnidirectional antennas. During the data collection, we maintain a sampling rate of 1000 Hz, and the devices operate at the 5 GHz frequency. To conduct the experiments, we invite the participation of eight volunteers. Each volunteer is instructed to walk back and forth along every solid line marked within the grid, as shown in Fig. 5. This activity continues for a duration of 20 minutes, allowing us to gather 180 minutes of CSI data for each volunteer within each scenario.

3.2. Performance Evaluation

The evaluation of MulGaFi is conducted using datasets obtained from four different scenarios. In the evaluation, MulGaFi demonstrates an impressive average accuracy of 98.11% across all four scenarios. This performance underscores the effectiveness of fusing time-frequency features with phase

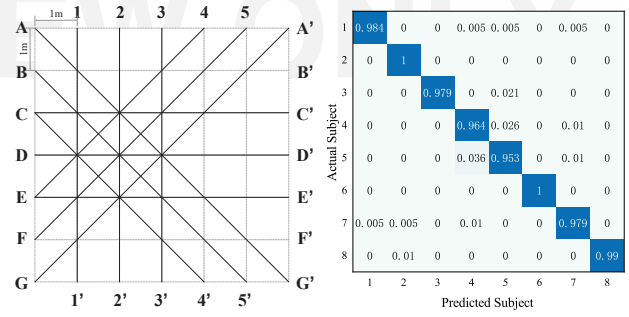


Fig. 5. walk routes of volunteers. Fig. 6. The confusion matrix of gait recognition

System	time-frequency model	phase model	MulGaFi w/o SEBlock	WifiU	AGait	MulGaFi
Accuracy	92.84	90.36	96.74	90.82	95.44	98.11

Table 2. Accuracy (%) of MulGaFi and Comparisons.

features. To provide further insights into the gait recognition results, the confusion matrix is presented in Fig. 6. Within this matrix, the diagonal entries indicate the true positive rate for recognizing different subjects.

Additionally, we compare MulGaFi with time-frequency model (utilizing only time-frequency feature extractor), phase model (utilizing only phase feature extractor), our model without SEBlock, the typical model WifiU, and the AGait model [13] which has datasets similar to ours. As shown in Table 2, MulGaFi demonstrates superior performance compared to our base models (time-frequency model, phase model and our model without SEBlock), highlighting the effectiveness of our well-structured fusion approach that fuses CSI spectrogram features and phase features. Furthermore, MulGaFi outperforms WifiU and AGait, underscoring its excellence in human gait recognition. Besides, we conduct an experiment to determine the principal component retention of PCA. we find that the first 12 principal components can effectively explain approximately 95% of the CSI amplitude data. Retaining these 12 principal components nearly achieves optimal performance, further ensuring performance while reducing time complexity of our system.

4. CONCLUSION

In this paper, we introduce MulGaFi, a novel multimodal-based gait recognition system using commodity Wi-Fi devices. MulGaFi operates by extracting and fusing gait-related features from two modalities: the CSI spectrogram and the CSI phase, which enhances the gait recognition performance. We conduct experiments in four distinct indoor settings with the participation of eight volunteers. The results demonstrate the remarkable performance of MulGaFi, with an average accuracy of 98.11%, highlighting the benefits of multimodal sensing in the field of gait recognition.

5. REFERENCES

- [1] Thiago Teixeira, Deokwoo Jung, Gershon Dublon, and Andreas Savvides, "Pem-id: Identifying people by gait-matching using cameras and wearable accelerometers," in *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. IEEE, 2009, pp. 1–8.
- [2] Toby HW Lam, King Hong Cheung, and James NK Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern recognition*, vol. 44, no. 4, pp. 973–987, 2011.
- [3] Weijun Tao, Tao Liu, Rencheng Zheng, and Hutian Feng, "Gait analysis using wearable sensors," *Sensors*, vol. 12, no. 2, pp. 2255–2283, 2012.
- [4] Robert J Orr and Gregory D Abowd, "The smart floor: A mechanism for natural user identification and tracking," in *CHI'00 extended abstracts on Human factors in computing systems*, 2000, pp. 275–276.
- [5] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Costas Spanos, "Wifi-based human identification via convex tensor shapelet learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.
- [6] Jing He and Wei Yang, "Imep: Device-free multiplayer step counting with wifi signals," *IEEE Transactions on Mobile Computing*, 2022.
- [7] Sheng Chen, Wei Yang, Yang Xu, Yangyang Geng, Bangzhou Xin, and Liusheng Huang, "Afall: Wi-fi-based device-free fall detection system using spatial angle of arrival," *IEEE Transactions on Mobile Computing*, 2022.
- [8] Zhengyang Wang, Sheng Chen, Wei Yang, and Yang Xu, "Environment-independent wi-fi human activity recognition with adversarial network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3330–3334.
- [9] Jin Zhang, Zhuangzhuang Chen, Chengwen Luo, Bo Wei, Salil S Kanhere, and Jianqiang Li, "Meta-ganfi: Cross-domain unseen individual identification using wifi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–21, 2022.
- [10] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie, "Gaitfi: Robust device-free human identification via wifi and vision multimodal learning," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 625–636, 2022.
- [11] Chenshu Wu, Feng Zhang, Yuqian Hu, and KJ Ray Liu, "Gaitway: Monitoring and recognizing gait speed through the walls," *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2186–2199, 2020.
- [12] Wei Wang, Alex X Liu, and Muhammad Shahzad, "Gait recognition using wifi signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 363–373.
- [13] Yang Xu, Wei Yang, Min Chen, Sheng Chen, and Liusheng Huang, "Attention-based gait recognition and walking direction estimation in wi-fi networks," *IEEE Transactions on Mobile Computing*, 2020.
- [14] Chenshu Wu, Zheng Yang, Zimu Zhou, Kun Qian, Yunhao Liu, and Mingyan Liu, "Phaseu: Real-time los identification with wifi," in *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 2015, pp. 2038–2046.
- [15] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi, "Xmodal-id: Using wifi for through-wall person identification from candidate video footage," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–15.
- [16] Jinmeng Fan, Hao Zhou, Fengyu Zhou, Xiaoyan Wang, Zhi Liu, and Xiang-Yang Li, "Wivi: Wifi-video cross-modal fusion based multi-path gait recognition system," in *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*. IEEE, 2022, pp. 1–10.
- [17] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [18] Frederick N Fritsch and Ralph E Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.
- [19] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.
- [20] Laurie Davies and Ursula Gather, "The identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782–792, 1993.
- [21] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM computer communication review*, vol. 41, no. 1, pp. 53–53, 2011.