Application and Development of Behavior Recognition Methods Based on Wireless Sensing and Image Multimodal Fusion

Zihan Liu

{ liuzh1923@mails.jlu.edu.cn }

College of Electronic Science & Engineering, Jilin University, Changchun, 130021, China

Abstract. A groundbreaking solution in removing the shortcomings of single-mode technologies is the multimodal approach to wireless sensing and visualization data, with the essential effect of detecting human actions indoors through the use of a multimodal approach. Wireless sensing provides many benefits, like contactless working, light-interference resistance, and privacy protection, but it is vulnerable to multipath propagation. Visualization modality is characterized by high spatial resolution and works with light conditions and posing the probability of confidential information leaking. This paper reviews and analyzes the research advancements on behavior recognition through the systematic review of the foundation to integrate these two modalities, data set construction rationale, fusion architecture planning, optimization, and reliability improvement. Finally, the article mentions the possible issues in the actual implementation, namely the generalization of the environment and optimization of energy consumption, or smart homes and elder care. The work can be used both in theoretical research and in engineering practice in the area of multimodal behavior recognition.

Keywords: Wireless Sensing; Wi-Fi, Image, Multimodal, Behavior Recognition.

1 Introduction

Human Activity Recognition (HAR) technology has become very popular over the past years, with various applications being used in home care or elderly homes, smart homes, security surveillance, and in health care. The main methods that are currently used in HAR technologies are wearable sensor-based solutions, visual-based solutions, radar-based solutions, and Wi-Fi-based solutions [1]. Wearable sensor-based HAR allows identifying the identity using human physiological features. Nevertheless, these wearable sensors necessitate the existence of ongoing wearing or carrying, which causes forgetfulness and inconvenience problems. The most widely studied and used method is image recognition technology that takes advantage of the ubiquitous cameras, but there are issues like privacy protection and lighting conditions. Radar, as well as Wi-Fi technologies, are classified as wireless sensors. The achievement of wireless sensor technology to monitor and detect human activity in an event without the need of invasive

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364846

procedures is becoming more popular. Although radar technology is more precise and provides more spatial coverage than Wi-Fi, it needs the structural purchasing and incorporation of specific radar sensors, making their realization quite expensive. Wi-Fi, on the other hand, is an inseparable element of the internet infrastructure and hence is extensively incorporated into everyday life. Its cheapness and the fact that it is simple to establish a major strengths[2]. Single-mode applications are also limited to a great extent. An example of such is in the fact that, although image recognition technology does not contradict human visual principles and the extensive use of cameras creates opportunities, privacy and problems with lighting prevent its use in behavior recognition.

Thus, the paper continues the information available in the literature to examine multimodal fusion techniques utilizing both wireless sensing and image recognition to identify behaviors and defines the way the technique is used in the future. It will overcome the bottlenecks of single-modal recognition technologies by combining the merits of both modalities. This paper will first focus on the low-cost and sustainable commercially deployable hardware technologies not considering those that incorporate radar-based special RF equipment and wearable sensors with images. Alternatively, it dwells on the behavior recognition methods in the indoor environment and integrates Wi-Fi sensing with multimodal images. This strategy preconditions the commercial implementation of new technologies in the future and increases the utility. Additionally, this study can find extensive use in various real-life applications such as home-based geriatric or elder care, residential homes, health care rehabilitation and security monitoring.

2 Core Technologies and Research Status

2.1 Research on Wireless Sensing Technology

Goods that can be found on the market are commercial Wi-Fi devices which are mainstream hardware used to acquire signals. This can be done via firmware updates through which these devices are able to retrieve granular CSI information, such as essential parameters, such as subcarrier amplitude, phase, and Doppler frequency shift (DFS) at very low cost [3]. But RAWS fine-grained CSI is prone to noise effects and needs multi-stage preprocessing in order to reduce the noise effects. Popular methods are the use of Hampel filters to eliminate outliers [4], Savitzky-Golay (SG) filters to eliminate high-frequency noise [5], and other studies have utilized wavelet transforms and principal component analysis (PCA) to reduce dimensions and maximize features [6].

The use of human activities has two main ways of impacting the signals to wireless networks. The mechanism first associated is associated with physical motion, which changes the path of signal propagation length and alters the phase parameter. The second process is associated with the speed and direction of movement and occurs as variations in the frequency parameters of signals that, in turn, cause a change in the DFS parameters. Deep learning algorithms are able to learn behavioral features of CSI time-series data, including the ability to convert one-dimensional arrays of CSI to two-dimensional recurrent graphs in order to keep bu set temporal correlations [7].

2.2 Research on Image Modality Technology

The data acquisition using computers is mainly done by modern image processing methods using two categories of cameras, RGB and depth cameras. The fact that they can exactly depict space information, including the physique of people and the structures of bones, is their major benefit [8]. The preprocessing phase usually uses object detection techniques to produce human silhouette maps and remove the background noise [9]. The main basis of feature extraction is based on pre-trained models, and it is proven that the C3D model is the most effective when compared to others, which require processing video time series and the visual transformer could be useful in extracting global spatial correlations. The methods of images are however, limited greatly because of the environment. In conditions of low light, the detection can be low (below 46 percent) [10].

2.3 Multimodal Fusion Principles and Hierarchies

There are three significant heterogeneity challenges in two-modal fusion: data dimensionality divergence, disparate sampling rates, and an uneven distribution of information. Concerning the difference in the data dimensionality, Wi-Fi data is presented in the form of a time-series sequence, whereas the image data are presented in the form of a two-dimensional matrix. In terms of a problem of sampling rate mismatch, Wi-Fi has the capability of up to 200Hz of sampling rate, but images normally run between 15 and 30fps [6]. Information asymmetry is caused by the excessive noise in the Wi-Fi messages in the environment and the frequent repetition of information in image messages. The fusion process is also complicated by modal failures. As an example, the occidental will flat the picture modal failure, whereas the Wi-Fi information will be distorted by multi-path interference. The modal failures also present other issues in the fusion, like loss of images through occlusion or distortion of Wi-Fi data through multiple path interference.

The existing methods of fusion can be divided into two fundamental levels: feature-level fusion and decision-level fusion. The first is that feature-level fusion involves the straightforward combination of preprocessed bimodal features. In another example, the WiVi gait recognition system is a neural network that combines video-computed ideal CSI and measured CSI, and transforms it into semi-ideal CSI, which is aligned with measured CSI. Instead, it uses a two-branch attention model to encode spatiotemporal relationships of the phase and DFS features [6]. The approach is effective in saving information and also achieving high utilisation levels. Nevertheless, its ability to resist interference is quite low, the noise produced by one modality could easily spread to the fusion findings. Secondly, the decision level fusion necessitates getting prediction probabilities of each modality with the independent models, whereas the overall outcome is the production of the fusion module. In the example of Wi-Fi systems and visual multimodal learning systems, a four-layer deep neural network with the SoftMax as the results can be used to lightly adjust the decision weights by applying fusion or gating techniques [3]. This level is highly robust and information-loss is a disadvantage.

3 Research Progress and Status Analysis

3.1 Multimodal Dataset Construction

Datasets are the foundation of fusion research, and existing achievements have evolved from single-modal to multi-dimensional integration. The characteristics of core datasets are compared in Table 1.

Table 1. Performance Comparison of Multimodal Datasets

Dataset	Modality Combination	Number of Scenarios	Number of Participants	Number of Behavior Categories	Sample Size	Core Advantages
XRF V2	Wi-Fi + IMU + RGB Video	3	16	30	853 sequences	Multi-device IMU integration; automatic temporal annotation
MM-Fi	Wi-Fi + RGB Video	2	12	8	4800 groups	Cross-environment data coverage; supports generalization testing
WiVi-Gait	Wi-Fi + Depth Video	1	1	18	1080 sequences	Multipath annotation; includes Angle of Arrival (AoA)/Time of Flight (ToF) spatiotemporal parameters
Custom-built	Wi-Fi + RGB Video	1	2	3	600 groups	High synchronization accuracy; includes no-light environment samples

One of the best examples of a database is XRF V2, which has been able to realize the concomitancy of the Wi-Fi signal, multi-unit inertial measurement unit (IMU), and video data, as one of the first of its kind. The Wi-Fi signal is transmitted at 5.64Ghz and its transmission data is 200 packets per second. The sampling rate of IMU information sent out by wearable technologies like watches and headphones is 25-50 Hz. The video information is captured in 720p resolution and 15 frames/second. This data also uses text-based entry commands in order to mark the start and end of actions automatically, which practically minimizes the errors that are easily made when using manual annotations. Implicit of cross-scenario research MM-Fi database was specifically designed to support the generalization verification in similar studies by offering the test samples with different room layouts and furniture organization [11].

3.2 Fusion Architecture Design and Performance Comparison

The key design philosophy at feature-level fusion is called deep feature alignment and collaborative extraction. The Wi-CBR algorithm forms a two-channel self-attention unit to pick out phase and DFS entities. Phase is the variation in dynamic paths, and Doppler frequency shift is the motion velocity. Group attention determines major sets of features and these sets are optimized through a gating mechanism to lessen the information entropy. The accuracy of the cross-domain recognition was also increased by 18 percent when the Widar 3.0 dataset was considered as opposed to single-modality methods. As an example, the WiVi gait recognition system creates ideal CSI, which is generated out of video, and transforms it to semi-ideal CSI, which is made up of the features of the environment, using neural networks, and matches it with real CSI. This perceives 98.0% correctness in three individual group recognition. The feature-level fusion provides good storage of information and requires the modal synchronization accuracy, and is computationally complex [7].

Decision-level fusion focuses on modality-independent modeling and result-optimized fusion. The Wi-Fi branch uses hollow convolutions and channel-spatial attention to extract the feature, and the video branch uses the C3D model to make predictions. Finally, DNN-based decision fusion attains 12 percent higher accuracy than feature-based fusion in noisy settings. Wi-Fi and a visual multimodal system of learning directly concatenates both the output of the SoftMax on dual-modal, and decides by the use of a four-layer DNN. It has a 97.5% accuracy in regular lighting and a 95.83% accuracy when it is in the dark. Decision-level fusion is more robust but has the trade-off of information loss and accuracy ceiling to an extent that current methods usually achieve accuracies 3-5% less than those of feature-level fusion [3].

3.3 Lightweight and Robustness Optimization Strategies

In order to meet edge deployment demands, the LiteWiHAR system suggests an optimization that requires both compression of the stacked blocks of ConvNeXt v2-Atto [2,2,6,2] to compress to [1,1,3,1], reducing the channels by half, and the addition of the SimAM parameter-free attention block to extract global features. The model itself takes 1.83MB; its FLOPs are 78.04M, and the model has an accuracy of 99.82 on NTU-Fi HAR dataset [5]. The CaiT model is a two-stage design, using a self-attention design and a class-attention design, with parameters of class embedding updated only. It records 98.78% accuracy on UT-HAR, and its parameters are 89.6% smaller than those of ResNet18 [12].

Robust optimization deals with innovation of data augmentation and learning parsimony. MaskFi offers the mask modeling(MI2M) framework that is unsupervised by contrasting features across models cross-modally and reconstruction of missing information using mask reconstruction. The method is less costly, 80 percent less expensive than supervised learning in cluttered environments, and better, 9 percent more accurate, than annotation. The contrastive learning and diffusion model method produces quality Wi-Fi information through the use of diffusion models, and video geometric transformations improve the diversity of samples. This attains more than 90 percent accuracy on the MM-Fi data set. The modality switching mechanism modulates weights dynamically, by using brightness detection, which enables automatic weight gain in Wi-Fi modality in dark environments, which provides environmental adaptability [11].

4 Innovation Points and Technical Framework

4.1 Dynamic Hierarchical Adaptive Fusion Framework

Current levels of fixed fusion are unable to cope with dynamic situations. Fusion on the feature level is more accurate in cases where both modalities are of high quality, but performance is abysmal in cases where one of the modalities falters. Decision-level fusion is highly robust in nature, but the upper limit of accuracy is restricted by loss of information. Thus, a dynamically controlled fusion level switching architecture on a modality basis is needed. This model has three fundamental modules. The first one is that the Modality Reliability Assessment Module (calculates in real time) the signal-to-noise ratio (SNR) of the Wi-Fi modality and the score of object detection confidence of the image modality. These values are put in a normalized state to have reliability coefficients(0-1). The level of fusion is activated at the feature-level in cases where both modalities are more than 0.8 reliable; decision-level fusion is activated in cases where either of the modalities is less than 0.5, and a hybrid mode in cases of intermediate level values of reliability. The second is the dual-path feature processing module. The feature-level path utilizes SWACmixT hybrid Transformers (Shifted Window Attention + Convolution) to combine dual-modality features, with a tradeoff between local and global information. The decision-level model does not involve any modality modeling and gives out probability distributions. The third one is the dynamic weight fusion layer. Fusion weights are computed according to reliability coefficients: the weight of the fusion result at the feature level is the result of $\max(\alpha,\beta)$, while the decision-level result weight is $1-\max(\alpha,\beta)$, where α and β represent the reliability coefficients for Wi-Fi and image data, respectively.

In this arrangement of a dynamic modal quality fusion strategy that fuses the benefits of feature fusion and decision fusion, the accuracy of the behavior recognition in the complex scenarios is enhanced to some degree.

4.2 Cross-Domain Enhanced Contrast-Diffusion Fusion Model

Current models display strong performance loss in cross-environment (e.g. room rearrangement and user interference) conditions, and XRF V2 demonstrates 70 percent accuracy when doing zero-shot testing. The bottleneck is the difference in the inter-domain features distribution and lack of data [8]. Thus, there is a need to have an adaptive architecture which can dynamically switch as the environmental changes occur. This model has three main elements. The first is a cross-modal contrastive alignment unit that is used to build a contrastive learning space between Wi-Fi CSI and image features. In the hope of accomplishing similarity and dissimilarity between features of related behaviors in different modalities, it uses InfoNCE loss to learn domain-invariant features. Second is the diffusion enhancement unit bimodal. To overcome the problem of target domain data sparsity, it produces target domain Wi-Fi CSI sequences under the condition of high-quality source domain samples, based on diffusion models. Photos go through random geometric transformations to become improved, and a point of cross-domain sample store is built. Third is the attention-guided fusion head, which uses a Channel-Spatial Attention (CSA) block to compute dynamic weights of the improved bimodal features, which focuses on areas that are strongly related to behaviors.

This framework performs well in new situations and multi-user interference complex environments, where the current methods perform poorly in, and much better, and behavioral recognition accuracy is considerably improved.

4.3 Privacy-Enhanced Lightweight Federated Fusion Architecture

Image modalities carry the risk of privacy leakage, whereas centralized training has the problem of data silo. Current lightweight models do not have protection mechanisms against privacy, which means they are not applicable in sensitive matters such as healthcare [12]. Thus, an adaptive architecture that ensures privacy protection strategies should be created. This model is made up of three components. To begin with, there is an edge-side lightweight encoder at the end. The Wi-Fi branch uses the deep separable convolutional compression of ConvNeXt, where the image branch uses the MobileNetV3, where model compression is achieved through the use of federated distillation. Two, the feature aggregation layer is federated with edge devices, transmitting only modal feature representation as opposed to raw data. The server combines features with a federated averaging algorithm and adds Gaussian noise through mechanisms of differentially private (DP) controls and keeping the privacy budget ε less than 1.0. Third is the central fusion decoder, which is an undertaking of making decision-level fusion utilizing amalgamated features. Multi-task learning or behavior recognition and privacy-preserving loss. The model is optimized in order to balance privacy, security and recognition accuracy.

The scope of the use of this architecture is diverse as it is applicable in different situations involving the protection of the privacy of an improved level, like in medical rescue and home-based care of the elderly. It has the maximum behavior recognition accuracy, which is convincingly better in terms of privacy.

5 Existing Challenges and Future Application Prospects

5.1 Core Technical Challenges

There are three fundamental challenges in terms of technology. First, a lack of cross-environment generalization: currently, the models are being trained within particular settings, which causes their performance to decrease severely in new environments, e.g. different room layouts, grease interference in kitchens, through multipath reflections in bathrooms, or multi-user interference, and no general, effective solution to the problem has been created. Second, poor synchronous intermodal accuracy. The error based on the sampling of Wi-Fi and image data leads to the resulting cases of time misalignment. In the existing techniques, synchronization accuracy reaches the 100ms range, which does not satisfy the requirements of fine-grained behavior recognition. Third, challenges in balancing both the energy usage and performance: Edge devices lack sufficient computing capabilities. Although there are current lightweight models that lower energy consumption, it also decreases the accuracy of the fine-grained behavior recognition by 5-8 percent [12].

5.2 Future Development Prospects

In terms of future development, there are three major technological directions that could be considered in innovation. To start with, multimodal large-model pre-training. Learn basic models with massive data, such as XRF V2, improve generalisation with the pre-training-scenario fine-tuning paradigm, and learn zero-shot behaviour recognition [8]. Second, edge-cloud collaborative architecture. The edge supports lightweight feature-extraction and modality-switching, and the cloud supports cross-domain model-optimization and multi-user collaborative computing with a balance between the energy consumption and performance.

Third, software-hardware co-optimization. Create low-power Wi-Fi sensing chips, which include CSI preprocessing units, with special AI acceleration units, such as TPUs, to improve the performance of edge inference.

6 Conclusion

The multimodal fusion of wireless sensing and image is highly effective in overcoming the technical bottlenecks of single-modal behavior recognition on the basis of complementary properties and accomplishes a series of breakthroughs in the formation of databases, fusion architecture, and optimization of the performance. Nevertheless, cross-environment generalization, privacy protection, and multi-user collaboration are some of the challenges yet to be limited by the practical application.

The original designs of the variant of dynamic hierarchical adaptive fusion and cross-domain enhanced contrast-diffusion models presented in this paper provide new solutions on the dimensions such as fusion strategies, domain adaptation capacity, and privacy security. Future advances in multimodal large-model pre-training, edge-cloud collaborative optimization, and hardware-algorithm co-design will be able to be deployed on a large scale to home-based elderly care, smart homes, and healthcare. This shall drive the intelligent perception out of the realm of technically attainable to that of engineerable and practical.

References

[1] Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., Liu, Y.: Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR) 54(4), 1–40 (2021)

[2] Yang, Z., Zhang, Y., Zhang, Q.: Rethinking fall detection with Wi-Fi. IEEE Transactions on Mobile Computing 22(10), 6126–6143 (2023)

[3] Zou, H., Yang, J., Das, H.P., et al.: WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 426–433. IEEE, Long Beach (2019)

[4] Wang, Q., Yao, Q., Li, Y., Zhang, Y., Xu, C., et al.: Plap: CSI-based passive localization with amplitude and phase information using CNN and BGRU. Mobile Information Systems 2023 (2023)

[5] Diaz, G., Sobron, I., Eizmendi, I., Landa, I., Coyote, J., Velez, M.: Channel phase processing in wireless networks for human activity recognition. Internet of Things 24, 100960 (2023)

[6] Fan, J., Zhou, H., Zhou, F., et al.: WiVi: WiFi-video cross-modal fusion based multi-path gait recognition system. In: 2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS), pp. 1–10. IEEE, Oslo (2022)

[7] Behzad, K., Zandi, R., Motamedi, E., et al.: RoboMNIST: a multimodal dataset for multi-robot activity recognition using WiFi sensing, video, and audio. Scientific Data 12(1), 326 (2025)

[8] Lan, B., Li, P., Yin, J., et al.: XRF V2: a dataset for action summarization with Wi-Fi signals, and IMUs in phones, watches, earbuds, and glasses. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9(3), 1–41 (2025)

[9] Li, K.: Wi-Fi, visual, and hybrid positioning methods for mobile terminals in indoor dynamic scenarios. Hunan University, Changsha (2023)

[10] Luo, F., Khan, S., Jiang, B., Wu, K.: Vision transformers for human activity recognition using WiFi channel state information. IEEE Internet of Things Journal 11(17), 28111–28122 (2024)

[11] Yang, J., Tang, S., Xu, Y., et al.: MaskFi: unsupervised learning of WiFi and vision representations for multimodal human activity recognition. arXiv (2024)
[12] Liu, C., Liu, Y., Hao, Y., et al.: LiteWiHAR: a lightweight WiFi-based human activity recognition system. In: 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), pp. 1–5. IEEE, Singapore (2024)