Research on Human Motion Detection Based on WiFi and Vision Dual Modality and Future Prospects

Xinlin Li

School of Electronic Information Engineering, Shanghai DianJi University, Shanghai, 201306,China

BaCl22222@outlook.com

Abstract. Human motion recognition has important application value in smart home, medical monitoring and other fields. However, the visual technology of traditional unimodal technology depends on light and is easily affected by occlusion. At the same time, wireless sensing technology is not sensitive enough to subtle movements, and WiFi signals cannot be effectively recognized. Therefore, this study proposes a dual-modal fusion scheme based on WiFi channel state information (CSI) and visual skeleton key points to improve the robustness of the system through feature-level complementarity. This paper first verifies the feasibility of WiFi unimodality based on the WiAR dataset, constructs a 180-dimensional feature engineering system, and uses a random forest model to achieve a test set accuracy of 72.49%. The experimental results show that this method has significant effects on medium-sized datasets. At the same time, the recognition rates of different actions are analyzed in detail, revealing the misjudgment pattern and its improvement direction. In addition, this study plans the integration scheme of the MPII visual branch and designs a dynamic weight allocation strategy to cope with environmental interference such as lighting changes. The research in this paper provides a new technical path for multimodal human motion recognition and has broad application prospects.

INTRODUCTION

As an important research direction in the field of computer vision and wireless sensing, human motion recognition has wide application value in smart homes, medical monitoring, security monitoring and other fields. With the rapid development of the Internet of Things and artificial intelligence technology, the demand for accurate recognition of human motion is growing[1,2]. Traditional motion recognition methods mainly rely on single-modal data, such as vision or wireless signals, and are divided into two categories: vision and wireless sensing. Visual methods are based on image or video data collected by cameras and classify motions by extracting key points of human skeletons or motion trajectories[3,4]. For example, tools such as OpenPose can efficiently extract human joint information, but their performance is significantly reduced in low-light or occluded scenes. Wireless sensing methods use the channel state information (CSI) of WiFi signals for motion recognition. Its advantage is that there is no need to deploy cameras, but WiFi signals have limited ability to capture subtle movements and are easily affected by environmental interference. Therefore, how to break through the limitations of single-modal technology and build a robust and adaptable motion recognition system has become an important challenge in current research.

In recent years, researchers have begun to explore multimodal fusion technology to improve the robustness of action recognition by combining the advantages of vision and wireless perception. For example, Zou et al. proposed a new human activity recognition scheme called WiVi, which uses multimodal machine learning of WiFi and vision to accurately and device-freely recognize common human activities[5]. The scheme has high recognition accuracy and robustness in occluded scenes, achieving an activity recognition accuracy of 97.5%. Furthermore, Bin Han, Lei Wang, Xinxin Lu and other researchers proposed a cross-modal meta-learning method based on model-agnostic meta-learning (MAML) for WiFi-based human activity recognition. The method assumes that the model can learn "learning methods" from thousands of different image classification tasks and apply them to WiFi-based human activity recognition[6]. By using only public image and WiFi signal datasets, the model trained by this method is able to recognize previously unseen activities using only 5 samples of each category, with an average accuracy of 88.5% in thousands of tests. In addition, Van-Anh Nguyen and Seong G. Kong developed a multimodal feature fusion model to address the problem of abnormal human behavior recognition under low-light conditions. By fusing feature maps extracted from visible light and thermal infrared videos, they effectively improved the performance of abnormal human behavior recognition in low-light scenes[7]. However, existing research still has deficiencies in feature extraction and modality fusion strategies, which need further optimization.

This paper aims to propose a dual-modal fusion scheme based on WiFi channel state information (CSI) and visual skeleton key points and improve the accuracy and robustness of action recognition through feature-level complementarity. First, this paper verifies the feasibility of WiFi single mode based on the WiAR dataset, constructs a 180-dimensional feature engineering system, and uses a random forest model for action classification; Secondly, This study plan to integrate the visual branch of the MPII Human Pose Dataset in the future and design a dynamic weight allocation strategy to achieve effective fusion of bimodal data; finally, the system performance is analyzed through experiments to explore the direction of improvement. The research in this paper provides a new technical path for multimodal action recognition, which has important theoretical significance and practical application value.

DATA AND METHODS

Data Source

The dataset contains 3521 samples from 3 volunteers. Each volunteer completed 16 types of actions (such as horizontal waving, high throwing, walking, etc.), and each action was collected 30 times. The data features include CSI amplitude, phase, time domain waveform and frequency domain energy distribution of 30 subcarriers × 3 receiving antennas × time series. The preprocessing method includes data cleaning (removing low-quality samples with signal-to-noise ratio <15dB), feature construction (decomposing complex numbers into real and imaginary parts and expanding them into 180-dimensional feature vectors) and noise suppression (sliding average filtering combined with wavelet threshold denoising) [8-9]. The complete code and preprocessing scripts can be obtained through https://github.com/linteresa/WiAR.

The MPII dataset is extracted from videos and contains 25,000 images covering 410 daily activities. The annotation information includes 16 joint point coordinates and 3D posture information. The data can be obtained by applying for https://www.mpi-inf.mpg.de/departments/ computer-vision-and-machine-learning/ software -and-datasets/mpii-human-pose-dataset. The image resolution is 1920×1080 pixels on average, and the annotation attributes include 2D/3D joint coordinates, action category labels, and occlusion status[10-11]. The preprocessing plan includes extracting joint point heat maps using the OpenPose model and calculating the temporal encoding of joint motion speed and acceleration based on adjacent frame displacements. The visual advantages of MPII data can make up for the lack of spatial resolution of WiFi signals, and feature-level fusion is achieved through timestamp synchronization and spatial mapping. The use of data must comply with the official agreement.

Method

This study plans to use dual-modal fusion technology, combining WiFi channel state information (CSI) and visual skeleton key points, to improve the robustness and accuracy of human motion recognition through feature-level complementarity. Based on CSI data, the WiFi modality constructs a 180-dimensional feature vector through complex decomposition, dimensional expansion and dual-channel processing to capture the subtle effects of human motion on wireless signals, and uses a random forest model for classification; the visual modality plans to use the MPII dataset to extract 16 joint point coordinates through OpenPose, and enhance the motion features based on temporal coding (such as joint displacement speed) to provide accurate spatial information. The dual-modal fusion designs a dynamic weight allocation strategy to adaptively adjust the weights of WiFi and visual modalities according to environmental conditions (such as light intensity)[12]. When the light is sufficient, the vision is the main mode, and when the light is low, the WiFi is the main mode, which improves the adaptability of the system in complex environments. The characteristics of this method are the originality of feature engineering, the efficiency of the model and the flexibility of the fusion strategy, which provides a new technical path for multimodal human motion recognition.

In the parameter configuration of the random forest model, the out-of-bag error (OOB) is an important evaluation indicator. The OOB error estimates the generalization error of the model by using unsampled data (i.e., out-of-bag samples) during the training process of each decision tree, avoiding the need for an additional validation set. By analyzing the convergence curve of the OOB error, it is possible to determine whether the model has achieved stable performance during the training process. Experiments show that as the number of decision trees increases, the OOB error gradually decreases and tends to stabilize. In this study, when the number of decision trees reaches 100, the OOB error tends to stabilize and reaches the optimal value, so 100 is selected as the number of decision trees.

In the experimental table, the parameter selection basis can be further refined to more clearly reflect the analysis process of OOB error. The modified table is as follows：

RESULTS ANALYSIS

TABLE 1 Random Forest Configuration

parameter	value	Selection basis
Number of decision trees	100	Out-of-bag error (OOB) convergence curve
Maximum characteristic ratio	0.156	Derived from the TreeBagger function
Split criteria	Gini Index	Classification task standard configuration

The parameter configuration of the random forest model is shown in Table 1. The number of decision trees is set to 100. This choice is based on the analysis of the out-of-bag error (OOB) convergence curve to ensure that the model can achieve stable performance during training. The maximum feature ratio is 0.1567, which is obtained through the TreeBagger function, aiming to balance the generalization ability and classification performance of the model. The splitting criterion uses the Gini index, which is a standard configuration in classification tasks and can effectively measure the purity of feature splitting, thereby improving the classification accuracy of the model. These parameters have been rigorously verified experimentally to ensure that the model has good performance on medium-sized datasets (Table 2).

TABLE 2 Analysis of action error patterns

Action Type	Success rate	False positive rate	Most misjudged actions
bend	73.0%	27.0%	hand clap
draw tick	79.5%	20.5%	two hands wave
draw x	74.0%	26.0%	draw tick
drink water	75.2%	24.8%	phone call
forward kick	49.7%	50.3%	hand clap
hand clap	69.8%	30.2%	bend
high arm wave	68.6%	31.4%	horizontal arm wave
high throw	75.7%	24.3%	high arm wave
Horizontal arm wave	76.3%	23.7%	high arm wave
phone call	84.2%	15.8%	toss paper
side kick	55.0%	45.0%	hand clap
sit down	83.8%	16,2%	walk
squat	69.9%	30.1%	two hands wave
toss paper	74.1%	25.9%	bend
two hands wave	76.9%	23.1%	horizontal arm wave
walk	77.7%	22.3%	drink water

Table 3 shows the recognition success rate, misjudgment rate and most misjudged action types of various actions. Overall, the recognition success rate of the model varies significantly among different actions, among which "phone call" has the highest success rate, reaching 84.2%, while "forward kick" has the lowest success rate, only 49.7%. In terms of misjudgment rate, "forward kick" and "side kick" have the highest misjudgment rates, 50.3% and 45.0% respectively, and both are most often misjudged as "hand clap". In addition, there is a high confusion rate between some actions, such as the misjudgment rate between "draw x" and "draw tick" is 26.0%, indicating that these actions have a high similarity in the feature space. From the perspective of misjudgment patterns, confusion between hand actions (such as "hand clap") and lower limb actions (such as "forward kick") is more common, which may be due to the insufficient sensitivity of WiFi signals to lower limb actions. Overall, this table reflects the performance of the model on different action categories and provides an important reference for subsequent optimization.

TABLE 3 Performance

Evaluation Metrics	Training set	Test Set
Accuracy	89.2%	72.49%

The experimental results show that the accuracy of the WiFi unimodal model on the training set is 89.2%, while the accuracy on the test set is 72.49%. However, the accuracy of the test set is significantly lower than that of the training set, indicating that the model has a certain degree of overfitting, which may be due to insufficient noise processing in feature engineering or uneven data distribution. Further analysis found that WiFi signals are not sensitive enough to lower limb movements, especially dynamic kicking movements (such as "forward kick" and "side kick") are easily confused with hand movements (such as "hand clap"). In addition, the confusion rate between "draw x" and "draw tick" is also high (26.0%), which may be due to the strong feature similarity of these two gestures in wireless signals.

There are some limitations in the current research. For example, the joint time-frequency information is lost in the process of complex number decomposition, which may affect the recognition ability of complex actions; WiFi signals are susceptible to multipath interference, especially in an environment with many metal objects, the signal attenuation is serious, resulting in a decrease in classification performance; in addition, the current experiment only uses data from three volunteers, and the sample diversity is insufficient, which may affect the generalization ability of the model. To address these problems, The study can introduce a joint time-frequency analysis method (such as wavelet transform or short-time Fourier transform) to extract richer action features, design a multipath interference suppression algorithm to improve environmental adaptability and expand the data set size to increase sample diversity to enhance the generalization ability of the model. Future research will combine the MPII visual data set to further improve system performance through dual-modal fusion and provide a more robust solution for human action recognition.

CONCLUSION

This study explored a new method for human motion detection based on WiFi and visual dual-modal technology. By fusing WiFi channel state information (CSI) with visual skeleton key points planned to be added in the future, a feature-level complementary scheme is proposed to improve the robustness and accuracy of the system. The study first completed the feasibility verification of WiFi single modality based on the WiAR dataset, constructed a 180-dimensional feature engineering system, and used a random forest model for classification, with a test set accuracy of 72.49%. The experimental results show that the WiFi modality has a certain classification ability on medium-sized datasets, but there is an overfitting phenomenon, and the sensitivity to lower limb movements is insufficient, and the misjudgment rate of some movements is high.

The study found that WiFi signals are stable in low light or occluded conditions, but have limited ability to recognize fine movements, especially large dynamic movements (such as kicking) that are easily confused with other movements. In addition, the joint time-frequency information is lost during the complex decomposition process, which may affect the representation ability of features. Future research will combine the MPII visual dataset, extract skeleton key points through OpenPose, and design a dynamic weight allocation strategy to achieve adaptive fusion of WiFi and visual modalities to further improve system performance.

The innovation of this study is that it proposes a technical path for dual-modal fusion, which provides a new solution for human action recognition. Its significance lies in overcoming the limitations of traditional single-modal technology and providing more robust technical support for applications in smart homes, medical monitoring and other fields. In the future, the study can further optimize the feature extraction method, expand the data set size, and explore more efficient fusion strategies to promote the actual implementation and widespread application of this technology.

REFERENCES

1. L. Guo, L. Wang, J. Liu, et al., A Survey on Motion Detection Using WiFi Signals, Proceedings of the 12th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), 202–206, (2016).
2. D. Halperin, W. Hu, A. Sheth, et al., Tool Release: Gathering 802.11n Traces with Channel State Information, ACM SIGCOMM Computer Communication Review, (2011).
3. M. Sultana and S. K. Jung, Illumination invariant foreground object segmentation using ForeGANs, arXiv preprint arXiv:1902.03120, (2019)

4. C. Pham, L. Nguyen, A. Nguyen, N. Nguyen, and V. T. Nguyen, Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks, Multimedia Tools and Applications, 80(19), 28919–28940, (2021)
5. H. Zou, J. Yang, H. Prasanna Das, et al., WiFi and Vision Multimodal Learning for Accurate and Robust Device-Free Human Activity Recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, (2019).
6. B. Han, L. Wang, X. Lu, J. Meng, and Z. Zhou, Cross-Modal Meta-Learning for WiFi-Based Human Activity Recognition, Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, (2023).
7. V. A. Nguyen and S. G. Kong, Multimodal Feature Fusion for Illumination-Invariant Recognition of Abnormal Human Behaviors, Information Fusion, 100, 101949, (2023).
8. L. Guo, L. Wang, J. Liu, et al., HuAc: Human Activity Recognition Using Crowdsourced WiFi Signals and Skeleton Data, Wireless Communications and Mobile Computing, (2018).
9. Linteresa, WiAR: WiFi-based activity recognition dataset, GitHub repository, (2019). Available: https://github.com/linteresa/WiAR. Accessed: Feb. 23, 2025.
10. M. Andriluka, L. Pishchulin, P. Gehler, et al., 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, CVPR, (2014).
11. Max Planck Institute for Informatics, MPII Human Pose Dataset, Dataset, (2014). Available: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/software-and-datasets/mpii-human-pose-dataset. Accessed: Feb. 26, 2025.
12. C. Pham, L. Nguyen, A. Nguyen, N. Nguyen, and V. T. Nguyen, Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks, Multimedia Tools and Applications, 80(19), 28919–28940, (2021)