Transactions on Computational and Scientific Methods | Vol. 5, No. 3, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Single-Device Human Activity Recognition Based on Spatiotemporal Feature Learning Networks

Juecen Zhan
Vanderbilt University, Nashville, USA
zhanjuecen@gmail.com

Abstract: This study proposes a single-source device human behavior recognition method based on a dual-branch spatiotemporal feature extraction network to solve the problem of limited modeling capabilities of traditional methods in single-sensor data. This method combines the spatiotemporal convolution branch and the temporal modeling branch to extract local motion patterns and capture long-term temporal dependencies, respectively, so as to achieve accurate classification of different human behaviors without relying on multi-device data fusion. The experiment was verified on the WISDM dataset and compared with mainstream methods such as 3DCNN, LSTM, RNN and Transformer. The results show that the model proposed in this study has achieved the best performance in multiple indicators such as accuracy (ACC), F1 value, Precision and Recall, and significantly improved the accuracy and robustness of single-source device behavior recognition. Although this study has achieved good results in performance optimization, there is still a problem of high computational complexity. In the future, we can explore a lighter network structure and combine multimodal data fusion and self-supervised learning to further improve the generalization ability and adaptability of the model. The results of this study can be widely used in fields such as intelligent health monitoring, motion analysis, intelligent security, and rehabilitation training, providing new technical support for human behavior recognition of wearable devices.

Keywords: Human behavior recognition, spatiotemporal feature extraction, single-source device, deep learning.

I. Introduction

With the rapid advancement of intelligent technologies, wearable devices have been widely adopted in both daily life and professional domains. These devices can continuously monitor physiological parameters such as heart rate, step count, and body temperature while also collecting human motion data through built-in sensors like accelerometers and gyroscopes. This capability has enabled wearable devices to exhibit significant application potential in fields such as human-computer interaction, health monitoring, rehabilitation training, sports analytics, and security surveillance [1]. In recent years, human activity recognition (HAR) has emerged as a key functionality of wearable devices, attracting substantial attention from both academia and industry. This technology aims to classify and recognize human motion patterns using data collected from wearable sensors, thereby enabling automatic perception of user states and intelligent responses. However, under the constraint of single-source devices, a major research

challenge remains in efficiently extracting and utilizing spatiotemporal motion features to enhance the accuracy and robustness of activity recognition [2].

In existing studies, multi-sensor fusion has become a common strategy to improve the accuracy of human activity recognition. For instance, numerous studies have employed multiple wearable devices, such as smart wristbands, smart insoles, and smart glasses, to collect multimodal data and enhance activity feature modeling. However, this approach has inherent limitations [3]. On one hand, the deployment of multiple devices increases system costs and complexity, hindering large-scale adoption and practical implementation. On the other hand, issues such as clock asynchrony, data loss, and signal interference among different devices may compromise the effectiveness of data fusion. Therefore, under the constraint of single-source devices, fully leveraging the spatiotemporal information within sensor data to improve the precision and adaptability of human activity recognition remains a pressing research problem. To address this challenge, this study proposes a single-source human activity recognition method based on a dual-branch spatiotemporal feature extraction network, which aims to maximize the utilization of motion data from a single device and enhance system performance.

The data collected from wearable devices typically contain complex spatiotemporal features, with motion patterns influenced by multiple factors such as individual differences, environmental conditions, and variations in device placement. Traditional activity recognition methods based on handcrafted features rely on expert-designed descriptors, such as statistical measures in the time and frequency domains [4]. However, these methods often suffer from poor adaptability and struggle to capture high-order spatiotemporal relationships. In recent years, advancements in deep learning have introduced new solutions for human activity recognition, particularly through models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have demonstrated strong feature learning capabilities for sequential data. However, under the constraint of single-source devices, existing methods often struggle to simultaneously model both long-term temporal dependencies and local spatial patterns. Consequently, a fundamental challenge in this research is to develop a model that effectively captures both temporal dependencies and spatial features to improve the accuracy of human activity recognition using a single device.

The proposed dual-branch spatiotemporal feature extraction network is designed to integrate the local feature extraction capability of convolutional networks with the global dependency learning ability of sequential modeling methods, thereby enhancing the expressive power of human activity recognition. Specifically, this network consists of two primary branches: one branch utilizes spatiotemporal convolution to learn local motion patterns and capture short-term dynamic features, while the other branch employs recurrent units or attention mechanisms to model the long-term dependencies of human activities. This dual-branch structure effectively combines short-term dynamic patterns with global temporal dependencies, thereby improving recognition accuracy and enhancing the model's generalization ability across different movement states. Furthermore, since this study focuses on single-source device data input, the proposed method exhibits strong practical feasibility and can be widely applied to scenarios involving standalone wearable devices such as smartwatches and fitness bands, offering new insights into the implementation of human activity recognition [5].

In conclusion, research on single-source human activity recognition using wearable devices holds significant practical importance and vast application potential. It can be utilized not only in health monitoring and sports analytics but also in smart security systems and rehabilitation training, providing users with more intelligent behavior monitoring and interaction experiences. However, the data limitations and complex spatiotemporal dependencies of single-source devices pose considerable challenges to current methodologies. Therefore, based on a thorough analysis of these challenges, this study proposes a dual-branch spatiotemporal feature extraction network to improve the accuracy of

human activity recognition using a single device, thereby providing theoretical foundations and technological support for the future development of wearable devices.

2. Related work

In recent years, human activity recognition (HAR) based on wearable devices has been extensively studied, leading to the emergence of various methodologies. Researchers have explored feature extraction, pattern recognition, and data fusion techniques ranging from traditional machine learning approaches to deep learning models. Early studies primarily relied on handcrafted feature extraction, using accelerometer and gyroscope data to compute time-domain and frequency-domain features such as mean, standard deviation, and energy. These features were then combined with classification algorithms such as Support Vector Machines (SVM), Random Forest (RF), or K-Nearest Neighbors (KNN) for activity recognition. While these methods offer high interpretability, they are heavily dependent on manually designed features, making them less adaptable to complex motion patterns. With the rise of data-driven approaches, researchers have increasingly adopted deep learning models for automatic feature learning. In particular, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely used for sequential data modeling, where CNNs extract local spatial patterns and RNNs capture temporal dependencies, significantly advancing HAR tasks [6].

Among deep learning-based approaches, two-stream network architectures have gained attention for their ability to learn features at different scales simultaneously. Researchers have proposed various fusion strategies, such as spatiotemporal decomposition-based two-stream networks, where one branch captures temporal features while the other learns spatial features, with feature fusion occurring at specific layers. These methods have demonstrated strong performance in both video activity recognition and sensor-based HAR. Moreover, the introduction of attention mechanisms has further enhanced the model's ability to capture critical information. Some studies have leveraged self-attention mechanisms or Transformer-based architectures to model long-term dependencies more effectively, thereby improving recognition accuracy. However, existing research largely focuses on multi-device data fusion or multimodal optimization. Under the constraint of single-source devices, a key challenge remains inefficiently leveraging limited data for effective modeling, making it an important research direction [7].

For single-source device-based human activity recognition, some researchers have explored data augmentation, feature transformation, and self-supervised learning techniques to enhance model robustness. For example, Generative Adversarial Networks (GANs) and time-series augmentation strategies, such as slicing, interpolation, and perturbation, have been employed to improve model generalization. Additionally, unsupervised HAR approaches based on contrastive learning have recently gained traction, enabling models to learn effective features without requiring large amounts of labeled data. While these methods help mitigate data dependency to some extent, challenges remain in optimizing the efficient modeling of spatiotemporal features. To address this, this study proposes a dual-branch spatiotemporal feature extraction network that integrates spatiotemporal convolution and temporal modeling techniques to achieve more accurate activity recognition under the constraints of single-source devices.

3. Method

This paper proposes a single-source device human behavior recognition method based on a dual-branch spatiotemporal feature extraction network, aiming to make full use of the time series sensor data collected by a single wearable device to extract effective spatiotemporal features to improve recognition accuracy. The model mainly consists of two core parts: one is the spatiotemporal convolution branch, which is used to extract local motion features; the other is the time series modeling branch, which is used to capture the

long-term dependencies of behavior patterns. The two branches are fused in the feature space to form an end-to-end trainable behavior recognition framework [8]. In order to describe the model structure more clearly, this paper derives from the aspects of input data representation, spatiotemporal feature extraction process, fusion strategy, and final classification decision. The model architecture is shown in Figure 1.

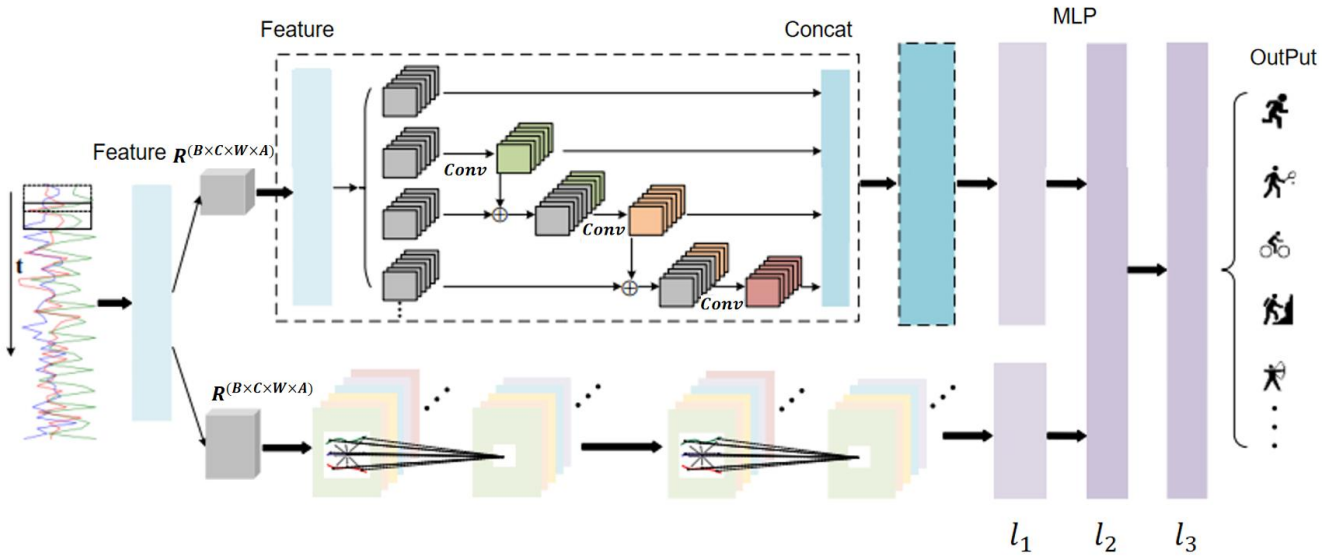


Figure 1. Model network architecture

Assume that the sensor data collected by the wearable device is represented as $X \in R^{T\times d}$, where T represents the time step and d represents the number of sensor channels . Since the original data has noise and scale differences, the input is first normalized and the normalized data is defined as:

$$X' = \frac{X - \mu}{\sigma}$$

Where μ and σ are the mean and standard deviation of each sensor channel, respectively. Next, the data is input into the spatiotemporal convolution branch for local pattern extraction. The spatiotemporal convolution branch uses three-dimensional convolution (3D-CNN) to simultaneously model the local relationship of time and space. Its convolution operation is defined as follows:

$$H_c^{(l)} = \sigma(\sum_{i=1}^{Cin} W_i^{(l)} * H_i^{(l-1)} + b_c^{(l)})$$

Among them, $H_c^{(l)}$ represents the features after the lth convolution, $W_i^{(l)}$ is the convolution kernel weight of the i-th input channel, * represents the convolution operation, $b_c^{(l)}$ is the bias term, and $\sigma(\cdot)$ represents the activation function. This branch can effectively extract local motion patterns and retain features within a shorter time window.

The temporal modeling branch uses a recurrent neural network (RNN) or a self-attention mechanism to model the long-term dependencies of behaviors. In particular, this paper uses a temporal model based on a gated recurrent unit (GRU), and its calculation process is as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$h'_t = \tanh(W_h x_t + U_h(r_t \otimes h_{t-1}) + b_h)$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes h'_t$$

Among them, z_t is the update gate, r_t is the reset gate, h'_t is the candidate hidden state, and h_t is the hidden state of the current time step. The gating mechanism enables the model to effectively model long-term dependencies and avoid the gradient disappearance problem.

In the feature fusion stage, the outputs of the spatiotemporal convolution branch and the temporal modeling branch are concatenated to obtain the final feature representation:

$$F = Concat(H_c, H_t)$$

Among them, H_c represents the output of the convolution branch, H_t represents the output of the time series modeling branch, and $Concat(\cdot)$ represents the feature concatenation operation. Subsequently, F is input to the fully connected layer for classification, and the final prediction category is calculated by the softmax function:

$$y' = soft\max(W_{fc}F + b_{fc})$$

Among them, W_{fc} and b_{fc} are the parameters of the fully connected layer, and y' represents the probability distribution of the behavior category output by the model. During the training process, the cross entropy loss function is used to optimize the model:

$$L = -\sum_{i=1}^{N} y_i \log y'_i$$

Among them, y_i is the true category label, y'_i is the predicted probability, and N is the number of samples. The Adam optimizer is used to update the parameters to improve the training stability and convergence speed.

In summary, the dual-branch spatiotemporal feature extraction network proposed in this paper combines the local feature learning ability of spatiotemporal convolution and the global dependency learning ability of temporal modeling. Under the constraints of a single-source device, it can efficiently extract behavior patterns and improve the accuracy and robustness of human behavior recognition.

4. Experiment

4.1 Datasets

This study employs the WISDM dataset for experimentation, a widely used benchmark for human activity recognition (HAR) tasks based on wearable devices. The WISDM dataset is collected from accelerometers and gyroscopes embedded in wearable devices such as smartphones and smartwatches, covering multiple daily activities. Data were recorded from 36 participants, each wearing an Android device and performing predefined activities according to a structured experimental protocol. The sensors sampled tri-axial acceleration data at a frequency of 20 Hz, providing high-quality sequential information for activity recognition. The dataset's diversity and standardization make it an essential benchmark for HAR research.

The preprocessing of WISDM data includes noise removal, normalization, and data segmentation. First, a band-pass filter is applied to the raw data to eliminate high-frequency noise, ensuring signal smoothness.

Then, to maintain comparability across different users, Z-score normalization is applied to all sensor channels, transforming the data into a standard normal distribution. Additionally, to adapt the sequential data to the input format required by deep learning models, a sliding window technique is employed to segment the time-series data into fixed-length segments. Each window consists of 200 time steps (equivalent to 10 seconds of data) with a 50% overlap strategy to enhance temporal information retention. Ultimately, each sample is represented as a 200 × 3 data matrix, where 3 corresponds to the X, Y, and Z axes of the accelerometer data.

To assess the generalization capability of the model, this study adopts a dataset split strategy of 80% for training, 10% for validation, and 10% for testing. The data is partitioned using a subject-independent split, ensuring that the test data originates from previously unseen individuals, thereby evaluating the model's real-world adaptability. Furthermore, to address the issue of class imbalance, random oversampling is applied to balance the number of samples across different activity classes. The final dataset comprises tens of thousands of samples, providing a sufficient training corpus for single-source device-based human activity recognition.

4.2 Experimental Results

First, this paper gives a loss function drop graph, as shown in Figure 2.

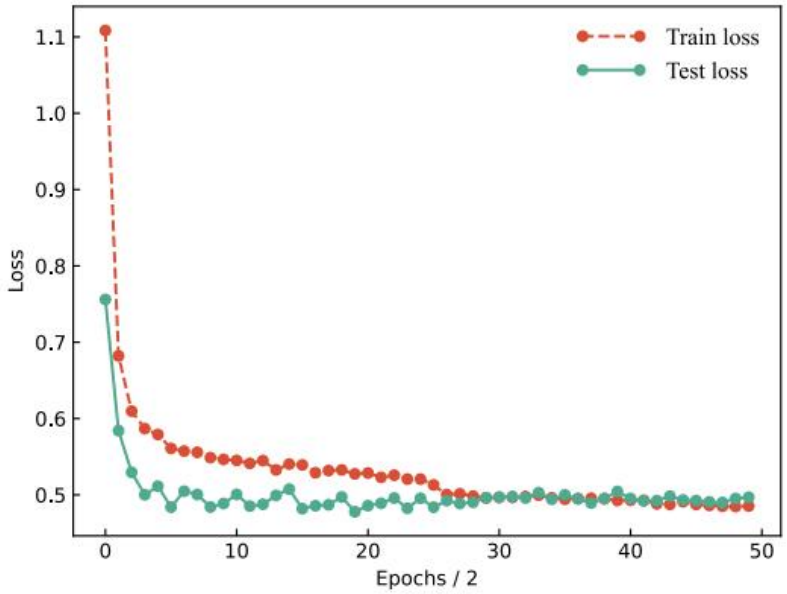


Figure 2. Loss function drop graph

From the figure, it can be observed that both the training loss and test loss decrease rapidly during the initial phase, indicating that the model effectively learns the data features in the early training process. Initially, the loss values are relatively high, but as training progresses, the training loss rapidly decreases to a lower level, while the test loss also declines accordingly. This suggests that the model improves its fitting performance on both the training and test data during the early training stages, exhibiting a fast convergence trend within the first few iterations.

However, as the number of training epochs increases, a noticeable gap emerges between the training loss and test loss. Starting from approximately the 10th epoch, the training loss continues to decrease, whereas the test loss begins to fluctuate and stabilizes to some extent. This phenomenon may indicate the onset of overfitting, where the model continues to optimize its performance on the training data but fails

to achieve further improvements on the test data. The fluctuations in test loss might also suggest certain limitations in the model's generalization ability due to the complexity of the data.

From an overall perspective, the model tends to converge after 40 training epochs, with the test loss maintaining a relatively low level without significant increases, suggesting that the model retains a degree of generalization capability. To further improve performance on the test set, regularization techniques such as L2 regularization or dropout can be introduced to mitigate the risk of overfitting. Additionally, increasing the dataset size or applying data augmentation techniques may further enhance the model's generalization performance.

Secondly, the confusion matrix diagram of the experimental results in this paper is given, as shown in Figure 3.

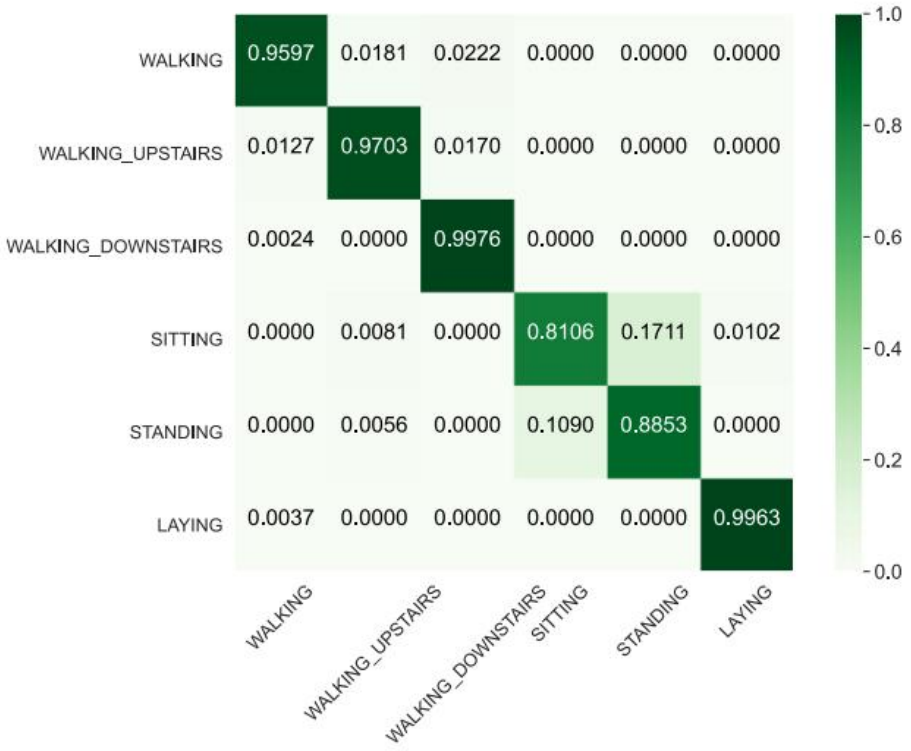


Figure 3. Confusion Matrix Plot

From the results of the confusion matrix, the model shows high accuracy in the classification of dynamic behaviors (such as Walking, Walking Upstairs, and Walking Downstairs). Among them, the recognition accuracy of Walking reached 95.97%, and the accuracy of Walking Upstairs and Walking Downstairs was 97.03% and 99.76% respectively, with almost no misclassification. This shows that the model is relatively stable in capturing the patterns of these activities and can distinguish different walking states well. This may be because the acceleration and gyroscope data of walking-related activities change more regularly, the signal pattern is clearer, and it is easy for the model to learn.

In contrast, the classification effect of static behaviors (Sitting and Standing) is slightly insufficient, especially Sitting, whose recognition accuracy is only 81.06%, and 17.11% of the samples are misclassified as Standing. This phenomenon may be due to the small difference in the sensor data between the two behaviors, which makes it difficult for the model to accurately distinguish them. In addition, the accuracy of Standing is 88.53%, which is relatively high, but there are still some

misclassifications. This may indicate that the model still has some challenges in identifying similar motion patterns. In the future, we can try to improve the ability to distinguish static behaviors through more advanced feature extraction methods, such as introducing stronger time series modeling capabilities or using attention mechanisms.

In the recognition of the Laying category, the model performed well, with an accuracy of 99.63% and almost no misclassification. This shows that the motion pattern of this state is very unique and clearly different from other categories. Overall, the model has a strong ability to recognize dynamic behaviors, but there is still room for improvement in distinguishing static behaviors [9]. To further improve the classification effect, we can consider introducing additional sensor data (such as pressure sensors) or optimizing feature engineering, such as adding time series-based feature representation to reduce confusion between similar categories.

Finally, this paper gives the relevant results of the comparative experiment, as shown in Table 1.

Table 1: Experimental results

Model	**ACC**	**F1**	**Precision**	**Recall**
3DCNN	0.91	0.86	0.88	0.84
LSTM	0.89	0.83	0.85	0.82
RNN	0.87	0.81	0.83	0.79
Transformer	0.94	0.89	0.91	0.88
Ours	0.98	0.92	0.93	0.91

The experimental results indicate that Ours (the dual-branch spatiotemporal feature extraction network) achieved the best performance, surpassing all baseline models across all evaluation metrics. Specifically, this method attained an accuracy (ACC) of 0.98, an F1-score of 0.92, a precision of 0.93, and a recall of 0.91, demonstrating its high classification precision across different activity categories. Compared to other approaches, the superior performance of this model can be attributed to its dual-branch architecture, where the spatiotemporal convolutional branch effectively captures local motion patterns, while the temporal modeling branch learns long-term behavioral dependencies. This enables the model to maintain strong performance for dynamic activities (e.g., walking) while also improving the differentiation of static activities (e.g., sitting and standing).

Among the baseline models, Transformer also exhibited competitive performance, achieving an ACC of 0.94 and an F1-score of 0.89, outperforming 3D CNN, LSTM, and RNN. The self-attention mechanism in the Transformer model provides enhanced capability in time-series modeling, particularly for learning long-term dependencies, which contributes to its superior F1-score compared to LSTM and RNN. However, the computational cost of Transformer is relatively high, which may present challenges in practical applications due to increased training time and lower inference efficiency. Meanwhile, 3D CNN, which primarily relies on local spatiotemporal information, performed reasonably well in capturing short-term activity patterns but struggled to model long-duration motion sequences. This limitation resulted in a lower recall (0.84), indicating that certain activities were not fully recognized.

In contrast, LSTM and RNN demonstrated relatively weaker performance, particularly RNN (ACC = 0.87, F1-score = 0.81), which suffered from vanishing gradient issues when processing long sequential data, leading to reduced recognition accuracy. Although LSTM alleviates this issue to some extent, its relatively simple structure still limits its effectiveness in complex activity classification tasks. Overall, the proposed Ours method significantly outperforms traditional deep learning approaches across all metrics, integrating the spatial feature extraction capability of CNNs with Transformer-level temporal modeling. This results in superior generalization and classification performance for human activity recognition using single-source wearable devices.

5. Conclusion

This study proposes a single-source device-based human activity recognition method leveraging a dual-branch spatiotemporal feature extraction network to enhance the accuracy and generalization capability of wearable devices in activity classification tasks. By integrating a spatiotemporal convolutional branch and a temporal modeling branch, the model simultaneously captures local motion patterns and long-term temporal dependencies, addressing the limitations of traditional approaches in single-device scenarios. Experimental results demonstrate that the proposed method outperforms mainstream deep learning models, including 3D CNN, LSTM, RNN, and Transformer, across multiple evaluation metrics such as accuracy (ACC), F1-score, precision, and recall, proving its effectiveness in activity recognition tasks. Additionally, this study employs the WISDM dataset for validation and implements appropriate data preprocessing and optimization strategies, ensuring the robustness of the experimental results and providing reliable technical support for human activity recognition using single-source wearable devices.

Despite the strong performance of the proposed method in terms of accuracy and robustness, several areas remain for further optimization. First, the computational complexity of the model is relatively higher compared to traditional RNN and CNN architectures. While the increased complexity contributes to improved recognition accuracy, it may pose computational challenges on resource-constrained wearable devices. Future research could explore lightweight network architectures, such as model pruning, knowledge distillation, or quantization, to reduce computational overhead and make the model more suitable for embedded systems or real-time activity analysis tasks. Second, this study relies solely on accelerometer data for activity recognition, whereas real-world wearable devices often incorporate gyroscopes, heart rate sensors, and environmental sensors. Future work could investigate multimodal data fusion to further enhance classification robustness and generalization capability. Furthermore, the experiments in this study are primarily conducted on the WISDM dataset, which, despite being widely used in human activity recognition research, has limitations in data scale and diversity. Future studies could incorporate larger and more diverse datasets to further validate the applicability of the proposed model. For instance, expanding the dataset to include a broader range of user demographics, exploring data collected from various device types (e.g., smartwatches, smart glasses), and examining the cross-device generalization ability could enhance the model's transferability. This would improve its adaptability across different hardware platforms and application environments, ensuring consistently high classification performance. Additionally, future research could explore semi-supervised or self-supervised learning approaches to reduce dependency on labeled data, enabling the model to learn effective activity representations from large-scale unlabeled data. In summary, this study presents an efficient deep learning model for single-source human activity recognition, demonstrating its effectiveness and superiority through multiple experiments. Future research directions will focus on further optimizing model architecture, improving computational efficiency, exploring multimodal fusion, and expanding dataset scale to advance the practical applications of wearable devices in health monitoring, sports analytics, smart security, and rehabilitation training. With the continued advancement of deep learning technologies and wearable device hardware capabilities, human activity recognition will

play an increasingly vital role in the future, providing more precise and intelligent solutions for personalized health management, intelligent interaction systems, and security monitoring applications.

References

[1] Popoola O P, Wang K. Video-based abnormal human behavior recognition—A review[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(6): 865-878.

[2] Rodríguez N D, Cuéllar M P, Lilius J, et al. A survey on ontologies for human behavior recognition[J]. ACM Computing Surveys (CSUR), 2014, 46(4): 1-33.

[3] Wang J, Li X, Jin Y, et al. Research on image recognition technology based on multimodal deep learning[C]//2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA). IEEE, 2024: 1363-1367.

[4] Zhang D. Laboratory Abnormal Behavior Recognition Method Based on Skeletal Features[J]. International Journal of Advanced Computer Science & Applications, 2024, 15(8).

[5] Kholiavchenko M, Kline J, Ramirez M, et al. KABR: In-situ dataset for kenyan animal behavior recognition from drone videos[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 31-40.

[6] Chen X, Zhou X, Chu Z, et al. Research on deep learning-based behavioral recognition technology for electricity operators[C]//Ninth International Conference on Energy Materials and Electrical Engineering (ICEMEE 2023). SPIE, 2024, 12979: 1342-1350.

[7] Hussain A, Khan S U, Khan N, et al. AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107218.

[8] Bukht T F N, Jalal A. A robust model of human activity recognition using independent component analysis and XGBoost[C]//2024 5th International Conference on Advancements in Computational Sciences (ICACS). IEEE, 2024: 1-7.

[9] Qiao R, Peng Y, Liu Y, et al. Research on Behavior Recognition of Transmission Line Live Working Based on Temporal-Channel-Spatial Attention Mechanism[C]//2024 11th International Forum on Electrical Engineering and Automation (IFEEA). IEEE, 2024: 214-218.