Expert Systems With Applications 238 (2024) 121978

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Image expression of time series data of wearable IMU sensor and fusion classification of gymnastics action

Yuliang Zhao [a,b], Fanghecong Dong [a], Tianang Sun [a], Zhongjie Ju [a], Le Yang [a], Peng Shan [a], Lianjiang Li [a], Xiaoyong Lv [a], Chao Lian [a,*]

[a] School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China
[b] Hebei Key Laboratory of Micro-Nano Precision Optical Sensing and Measurement Technology, Qinhuangdao, 066004, China

ARTICLE INFO

Keywords:
Human Activity Recognition (HAR)
Wearable motion sensors
Image encoding
Time series data
Gymnastics action recognition

ABSTRACT

Human activity recognition (HAR) based on multimodal wearable motion sensors is a valuable technology, and gymnastics action recognition is one of its important application scenarios. However, this field faces challenges such as excessive sensor node deployment, hindering effective utilization of time-series sensor data, and low accuracy in action recognition. To address these issues, this paper introduces a gymnastics action recognition system comprising a hardware setup and an action recognition framework that combines a time series data imaging method with a novel three-channel convolutional model ViTGS. Firstly, a cost-effective single-node sensor system equipped with a 6-axis IMU is devised to efficiently capture gymnastics action data. Secondly, three one-dimensional time series data image encoding methods, Gramian Angular Field (GAF), Recurrence Plot (RP), and Markov Transition Field (MTF), are employed to capture the temporal features, phase space features, and state transition features of the data respectively. Finally, the ViTGS model is proposed, which uses the ViT_B/16 network to extract the features of each channel image, and uses Gated Fusion Network (GFN) to realize the effective fusion of different channel features and uses SVM as the recognition layer to further improve the classification accuracy. The experimental results show that using single-node (left wrist) sensor data, an accuracy of 98.56% is obtained on the acceleration dataset, and an accuracy of 99.04% is obtained on the angular velocity dataset. The method of this paper is not only applicable to gymnastics action recognition but also can be generalized to other human action recognition fields.

1. Introduction

The technology of Human Activity Recognition (HAR) is a burgeoning and significant realm of research, and has been extensively implemented across a range of domains, such as medical healthcare, athletic conditioning, and psychological scrutiny (Han et al., 2022; Lian et al., 2021, 2020; Mim et al., 2023; Sha et al., 2020; Zhang, Li, Ge et al., 2022). Due to the continuous development of sensor technology, pattern recognition, and artificial intelligence, an increasing number of scientific research institutions have started to study HAR (Dua et al., 2021; Wang & Liu, 2020; Xu & Yuan, 2021). Among various human activities, gymnastics is globally recognized as an important sport due to its competitive and artistic nature. Gymnasts are required to complete various complex movements within a specified time frame based on musical accompaniment, and any gymnastics movement that does not meet the expected standards may hinder the athlete's performance or even cause physical injury (Wang & Hu, 2021). Therefore, the detection and evaluation of gymnastics movements still hold great importance.

HAR can be divided into two main directions: video-based HAR and wearable sensor-based HAR. Initially, video-based HAR commonly used low-cost and easy-to-use RGB cameras to record videos (Nafea et al., 2021). For instance, Lee and Ahn (2020) proposed a technique using a single RGB camera to recognize human behaviors, achieving a recognition rate of 70% on the NTU-RGBD dataset. However, conventional RGB cameras only record color information and cannot accurately capture depth information, which limits their ability to capture all necessary motion information in space (Yadav et al., 2021). Compared to RGB, RGB-D cameras can overcome this limitation by capturing both color and depth information, making them more effective for HAR. With the depth information captured by the RGB-D camera, video-based HAR can more easily distinguish similar movements that be difficult to differentiate in RGB videos. For instance, Siddiqi et al. (2021) developed an unsupervised segmentation algorithm that accurately segments the human body from video frames captured by a depth camera. They subsequently used a hidden Markov model (HMM) for

* Corresponding author.
E-mail addresses: zhaoyuliang@neuq.edu.cn (Y. Zhao), 2210326@stu.neu.edu.cn (C. Lian).

https://doi.org/10.1016/j.eswa.2023.121978
Received 9 June 2023; Received in revised form 25 September 2023; Accepted 1 October 2023
Available online 17 October 2023

Acronym

BiLSTM	Bidirectional Long Short-Term Memory
CART	Classification and Regression Trees
DCNN	Deep Convolutional Neural Network
FC	Fully Connected
GADF	Gram Difference Angle Field
GAF	Gramian Angular Field
GASF	Gram Sum Angle Field
GFN	Gated Fusion Network
HAR	Human Activity Recognition
HMM	Hidden Markov Model
IMU	Inertial Measurement Unit
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
MEMS	Micro-Electro-Mechanical Systems
MTF	Markov Transition Field
NB	Naive Bayes
QDA	Quadratic Discriminant Analysis
RP	Recurrence Plot
ViT_B/16	Vision Transformer Base 16 × 16
ViTGS	ViT_B/16-GFN-SVM
1D-CNN	One-Dimensional Convolutional Neural Network

training and testing. However, the high cost and complexity associated with using RGB-D cameras limit their widespread application in video-based HAR (Wang & Liu, 2020).

HAR based on wearable sensors is an increasingly important research direction (Huang et al., 2021; Mekruksavanich & Jitpattanakul, 2021; Sena et al., 2021; Wang, Cang et al., 2019). Compared with video-based method, sensor-based method has several advantages, such as low-cost experimentation, unconstrained data collection, and protection of subjects' privacy (Zhao et al., 2019). The process of sensor-based HAR is roughly as follows: first, various data on the target movement are collected using wearable sensors. For example, Niemelä et al. (2023) detects the posture of the participants using a waist accelerometer; second, the data is processed to extract key features about the target movement; finally, the extracted features are evaluated (Zhang et al., 2017). For instance, Sun et al. (2020) proposed a method for recognizing trunk movements based on inertial sensors. They classified 16 types of trunk movements using data collected by inertial sensors through quadratic discriminant analysis (QDA) and support vector machine (SVM).

However, in sensor-based HAR, representing and extracting features from one-dimensional time series sensor data is a challenging task. Prior to the application of deep learning in this field, HAR using wearable sensors relied on manually segmented features (Challa et al., 2022; Lu et al., 2017; Morales & Akopian, 2017). This method heavily relied on domain experts for data segmentation (Nweke et al., 2019). which was not only time-consuming but also had limited practical value (Wang, Chen et al., 2019). Additionally, manually extracted features are only a subset of the original data, and using them alone may lead to loss of raw information, thus hindering the generalizability of the method. These issues make it difficult to fully utilize the continuous time series data collected by wearable sensors.

With the continuous advancement of deep learning technology and its substantial achievements in classification problems (Zaitseva et al., 2023), researchers are increasingly utilizing this methodology in the field of HAR (Gholamiangonabadi et al., 2020; Ronao & Cho, 2016; Thu & Han, 2021; Zhang, Li, Zhang et al., 2022). By constructing multi-layer neural networks, it automatically detects a variety of features from raw data and learns both low-level and deeper-level features, eliminating the tedious manual segmentation process and improving the accuracy of HAR (Thirunavukarasu et al., 2022). This makes deep learning technology more advantageous over traditional machine learning techniques (Alawneh et al., 2021; Ramanujam et al., 2021; Shayesteh et al., 2023). For example, Yen et al. (2020) proposed a feature learning technique based on one-dimensional convolutional neural networks, which can recognize six types of daily human activities (walking, going up stairs, going down stairs, sitting, standing, and lying down). Similarly, Ramanujam et al. (2021) used CNN and gated recurrent units to automatically extract features and classify time series data.

However, deep learning still faces challenges in handling time series data (Challa et al., 2022). Despite the excellent performance of deep learning-based methods in HAR, several challenges remain in the analysis of time series data, including the presence of significant data bias and the complexity of extracting both global temporal features and local spatial features from activity data simultaneously (Challa et al., 2022; Jameer & Syed, 2023; Zhang et al., 2023). For example, Shi et al. (2023) used 1D-CNN and bidirectional LSTM networks to extract high-dimensional features and time series features from one-dimensional data collected by inertial sensors. The average recognition rate reached approximately 96%, suggesting potential for further improvement.

For the limitations of various previous HAR studies, we summarize as follows: According to Table 1, firstly, there is a concern about the high cost of data collection equipment. Video-based HAR often relies on expensive camera devices (Dhiman & Vishwakarma, 2019; Huang & Yan, 2020; Jalal et al., 2012; Lee & Ahn, 2020; Peng & Zhang, 2021; Siddiqi et al., 2021). Similarly, some sensor-based HAR methods utilize multiple sensors or expensive sensors, leading to cost escalations (Bhattacharya et al., 2022; Challa et al., 2022; Gupta, 2021; Lu et al., 2017; Mekruksavanich & Jitpattanakul, 2021; Sun et al., 2020). These factors limit the feasibility and widespread adoption of these approaches in practical applications. Secondly, some sensor-based HAR approaches adopt multi-node methods to collect data. This adds intricacy to the experiments, inconveniences the data collection and processing processes, and may potentially introduce privacy and security risks for participants (Bhattacharya et al., 2022; Challa et al., 2022; Dua et al., 2021; Haider et al., 2020; Han et al., 2022; Mim et al., 2023; Sun et al., 2020). In conclusion, the methods used for feature extraction limit the accuracy of these studies. The use of non-deep learning algorithms may result in a decrease in the accuracy of action recognition (Ahmed et al., 2020; Dhiman & Vishwakarma, 2019; Jalal et al., 2012; Lu et al., 2017; Sun et al., 2020). Even when using deep learning methods such as CNN for feature extraction, there are still difficulties in dealing with one-dimensional sensor data, which hinders the full potential of deep learning and ultimately affects the accuracy of HAR (Lee & Ahn, 2020; Sena et al., 2021; Thu & Han, 2021; Yen et al., 2020).

All the aforementioned methods focus on processing one-dimensional data, while a more effective approach is to transform the one-dimensional data into high-dimensional space and analyze it from multiple perspectives. Currently, re-representing the features of time series data as visual images has attracted the attention of researchers. Time series visualization can use supervised and unsupervised learning in computer vision to encode time series data into corresponding activity images, enabling deep learning techniques, particularly CNN, to perform image recognition. This approach allows machines to perform visual recognition and classification by learning the patterns and structures of visual images, replacing the direct analysis of one-dimensional data (Sarkar et al., 2023). For example, Qin et al. (2020) encoded each time series data of multiple sensors into dual-channel images for training. Furthermore, this method can be combined with deep learning to extract features and classify data more efficiently, thereby better utilizing the raw information in time series data. Sarkar et al. (2023) employed the continuous wavelet transform to encode

Table 1
Comparison of previously proposed HAR methods and our method.

Type	Literature	Sensor type	Node	Position	Method	Price	Accuracy
Video	Siddiqi et al. (2021)	Kinect	1	–	Unsupervised segmentation algorithm, Hidden Markov Models	200$	97.3%
	Lee and Ahn (2020)	RGB camera	1	–	3D skeleton, CNN	500$	70.0%
	Dhiman and Vishwakarma (2019)	Kinect	1	–	Average Energy Silhouette Images	200$	95.9%
	Jalal et al. (2012)	RGB-D camera	1	–	R transformation, Hidden Markov Models	600$	93.2%
	Huang and Yan (2020)	RGB Camera	1	–	Long Short-Term Memory (LSTM)	500$	82.8%
	Peng and Zhang (2021)	Kinect	1	–	Generating skeleton node graph	200$	90.0%
Sensor	Mekruksavanich and Jitpattanakul (2021)	IMU (Galaxy S II)	1	Waist	LSTM	600$	92.43%
	Yen et al. (2020)	IMU	1	Waist	1D CNN	30$	93.8%
	Challa et al. (2022)	IMU	3	Arm, chest, Ankle	LSTM, CNN	90$	94.3%
	Sun et al. (2020)	IMU	4	Spine	Quadratic discriminant analysis	120$	95.2%
	Ronao and Cho (2016)	IMU	1	Waist	CNN	30$	94.8%
	Han et al. (2022)	IMU	3	Arm, chest Ankle	CNN	90$	93.0%
	Mim et al. (2023)	IMU	3	Arm, chest Ankle	Gated Recurrent Unit-Inception (GRU-INC)	90$	90.3%
	Dua et al. (2021)	IMU	3	Arm, chest Ankle	CNN	90$	95.3%
	Sena et al. (2021)	IMU	1	Waist	DCNN	30$	90.8%
	Zhao et al. (2019)	IMU	2	Fingers	CNN	60$	94.8%
	Challa et al. (2022)	IMU	1	Waist	CNN, BiLSTM	30$	96.1%
	Lu et al. (2017)	IMU	3	Sacrum, arms	Unsupervised classification method (MCODE)	600$	86.0%
	Haider et al. (2020)	IMU, Mag, Bar	2	Waists	Super-bagging method	60$	84.2%
	Ahmed et al. (2020)	IMU	1	Waist	Feature selection, SVM	30$	96.8%
	Gupta (2021)	IMU (Galaxy S5, LG G)	2	Waist, wrist.	CNN-GRU	900$	96.5%
	Thu and Han (2021)	IMU	1	Waist	CNN, BiLSTM	30$	96.2%
	Bhattacharya et al. (2022)	IMU	3	Arm, chest ankle	CNN, LSTM	90$	97.5%
	Our work	IMU	1	Left wrist	Image encoding, ViTGS	30$	98.56%/99.04%

the time series data into multi-channel images; they further utilized a spatial attention-assisted CNN to extract features.

In conclusion, there are still several issues in wearable sensor-based HAR methods. Firstly, traditional approaches heavily rely on domain experts and require significant time and human resources for data segmentation. However, this method cannot guarantee that the extracted features meet the requirements of HAR, leading to a decrease in action recognition accuracy. Then, there are still challenges in using deep learning methods to process one-dimensional time series data. Existing algorithms fail to fully exploit the potential of deep learning in mining information and extracting comprehensive spatiotemporal features, thus, the full potential of deep learning in HAR has not been fully unleashed. Finally, using cameras or multiple sensors to collect human motion data increases costs and complicates experimental procedures. Therefore, while pursuing high-accuracy human activity recognition, it is essential to address the challenge of reducing experimental costs and workload.

To solve these problems, a novel framework for HAR has been proposed in this paper which combines temporal data visualization from sensors and a new three-channel convolutional model called ViTGS. Through this framework, action features can be automatically extracted and accurately classified. The primary contributions of this work are as follows:

(1) A hardware system has been developed for the collection of gymnastics movement data. This system employs single-node wearable sensors to collect acceleration and angular velocity data from various body regions during motion.

(2) The one-dimensional time series data is transformed into images using three image coding methods (GAF, RP, MTF) to reflect the data features from multiple perspectives. Among them, GAF captures the time domain features of the data, RP expresses the phase space features of the data, and MTF reflects the changes in the data state. The information from the time series data is then extracted from multiple dimensions.

(3) A novel three-channel convolutional network model, named ViTGS, is proposed to extract, fuse, and classify the transformed images.

- The model employs the ViT_B/16 network to extract image features, eliminating the need for manual dataset division and minimizing human factor interference.
- Subsequently, an efficient Gated Fusion Network (GFN) is employed to fuse the extracted features, thereby emphasizing the differences between action category features and enhancing the classification accuracy.
- Finally, the Full Connection and other machine learning algorithms are replaced by SVM to effectively classify the fused features. This substitution led to a significant improvement in the accuracy of the classification process.

The following sections of the paper are structured as follows: In Section 2, the acquisition and processing of data and the principles of the action recognition model are introduced. Detailed experimental analysis is presented in Section 3. Finally, this article is concluded in Section 4.

2. Method

The gymnastics action recognition system we proposed includes a hardware model and a three-channel feature fusion classification model, as shown in Fig. 1. Firstly, we obtain the time series data of acceleration and angular velocity of gymnastics actions through MEMS sensors. Next, we preprocess the collected data to eliminate noise. Finally, we encode the time series data into images and send them to ViTGS for action recognition.

2.1. Hardware system setup

The hardware components of the gymnastics action recognition system consist of a motion detection module, a main control module, a wireless communication module, and a personal computer module. Among them, the motion detection module is equipped with an MPU6500 chip, which has the ability to detect acceleration and angular velocity. The range of acceleration and angular velocity measured by the MPU6500 chip is set to ± 2 g $\sim \pm 16$ g and $\pm 250°/s \sim \pm 2000°/s$, respectively. In addition, the accelerometer and gyroscope of the MPU6500 chip measure three-axis data, with power consumption of 3.2 mA and 450 μA, respectively. The working temperature range

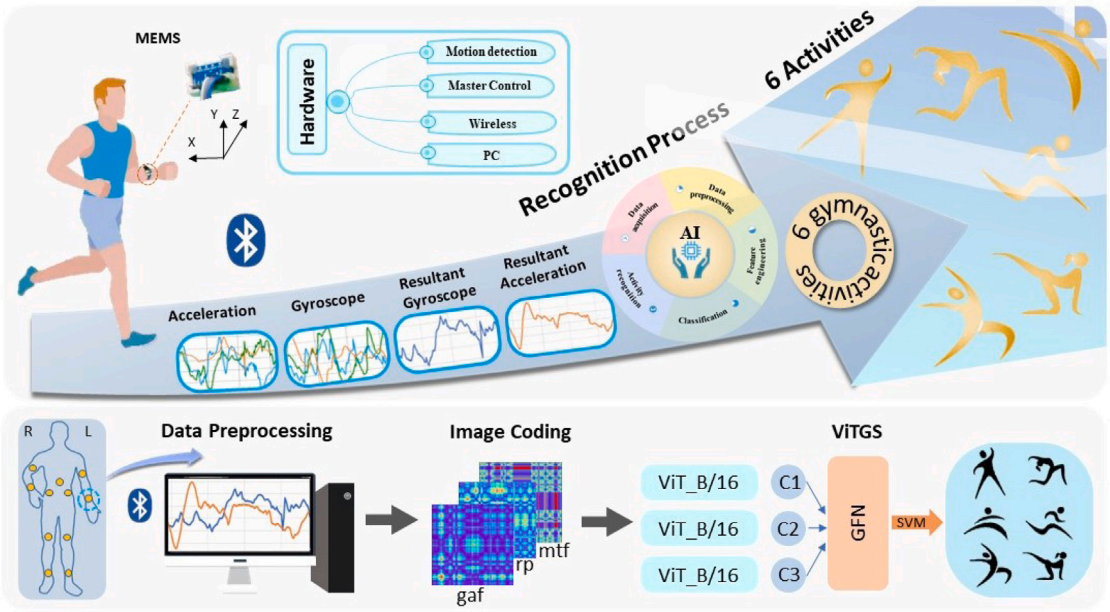

Fig. 1. Gymnastic movement recognition system that processes data and recognizes 6 different gymnastic movements.

is −40 – 85 °C, and the communication mode is I^2C. The primary function of the gymnastics action recognition system's control module is to receive and process data from the motion detection module in real-time. This module employs a 32-bit ARM microcontroller. The wireless communication module relies on ESP8266 WiFi, an economical WiFi SOC module capable of swiftly transmitting data to a personal computer.

2.2. Data processing

The experiment collected three-axis acceleration and three-axis angular velocity data from 11 positions on the human body, including the wrist, arm, hip, waist, leg, and foot, through MEMS sensors as shown in Fig. 1. The experiment included four male college students and three female college students between the ages of 20 to 26, all in good health. The height of male participants ranged from 170 cm to 178 cm, and their weight ranged from 66 kg to 80 kg. Female participants' height ranged from 158 cm to 160 cm, and their weight ranged from 52 kg to 55 kg. The Northeastern University Ethics Committee reviewed and approved the experiment. The subjects were instructed to perform six gymnastic movements: stretching (S), chest expansion (C), twisting (T), leg kick (L), body side (B), and full-body movement (F). The experiment consisted of five sessions, during which the seven participants repeated each movement eight times. Thus, for each data collection position, 280 sets of data were obtained for each type of action, resulting in a total of 1680 sets of data for the six types of actions.

Our specific data preprocessing steps are as follows:

(1) Sensor calibration: Due to external factors such as ambient temperature and noise, sensors inevitably have inherent drift, which may result in certain errors in the collected data and thus affect action recognition accuracy. In order to accurately evaluate the impact of sensor inherent errors on action recognition, we set the measuring range of the MPU6500 accelerometer sensor to ±16 g (corresponding to a sensitivity of 2048 LSB/g). We measured the static errors of the accelerometer on each axis using the following steps. The calculation formula for static errors is:

$$S = \frac{A - W}{W} \tag{1}$$

where S represents static error, A represents the actual value of the sensor, and W represents the measured value of the sensor.

The specific sensor calibration steps are as follows: the MPU6500 sensor is placed horizontally on the desktop, maintaining a static state. First, make sure that the X and Y axes of the sensor are parallel to the desktop, while the Z axis is perpendicular to the desktop and designated as the positive direction. At this time, the sensor system is only affected by gravity. The static error of the sensor can be obtained by measuring the difference between the three-axis acceleration and the ideal values (0 g for X and Y axes, 1 g for Z axis). Then, respectively set the positive direction of the X and Y axes to be vertically upward from the desktop to calculate the static error. The results show that the range of static acceleration error for all three axes is within ±0.03 g and does not change with the axis. When performing gymnastic movements, the measured values of three-axis acceleration range from −0.5 g to 2.2 g, −1 g to 2.8 g, and −1.8 g to 1 g, respectively. Compared to the actual measurement range, the range of static error accounts for only 1.6% to 2%, so the impact of static error on the accuracy of motion recognition can be considered negligible and can be ignored.

(2) Data collection: We use sensors to collect triaxial acceleration data and triaxial angular velocity data of gymnastics movements. The data sampling frequency is set to 50 Hz We combine the three-axis acceleration and three-axis angular velocity into total acceleration and total angular velocity, with the specific formula as follows:

$$R = \sqrt{X^2 + Y^2 + Z^2} \tag{2}$$

where R represents the total acceleration or total angular velocity, and X, Y, Z represent the data of the X-axis, Y-axis, and Z-axis.

(3) Data segmentation: In this study, we used a high-speed camera as an auxiliary device to validate the inertial information received by the sensor and to confirm it through manual segmentation during data processing. We selected the BASLER acA2000-165tm as our camera model because it meets our requirements for frame rate, resolution, and cost. This camera is capable of capturing rapidly moving objects in indoor motion scenes and presenting high-definition video footage. By simultaneously recording the data from the high-speed camera and the sensor, we obtained video data and sensor data with the same time axis. Based on the complete action timestamps in the video, we segmented all the sensor data.

(4) Data filtering: The data collection process is often disrupted by unintended trembling of the participants and other external factors. This affects the accuracy of gymnastic action recognition and evaluation. Interference signals are unavoidable in the raw data, but they can

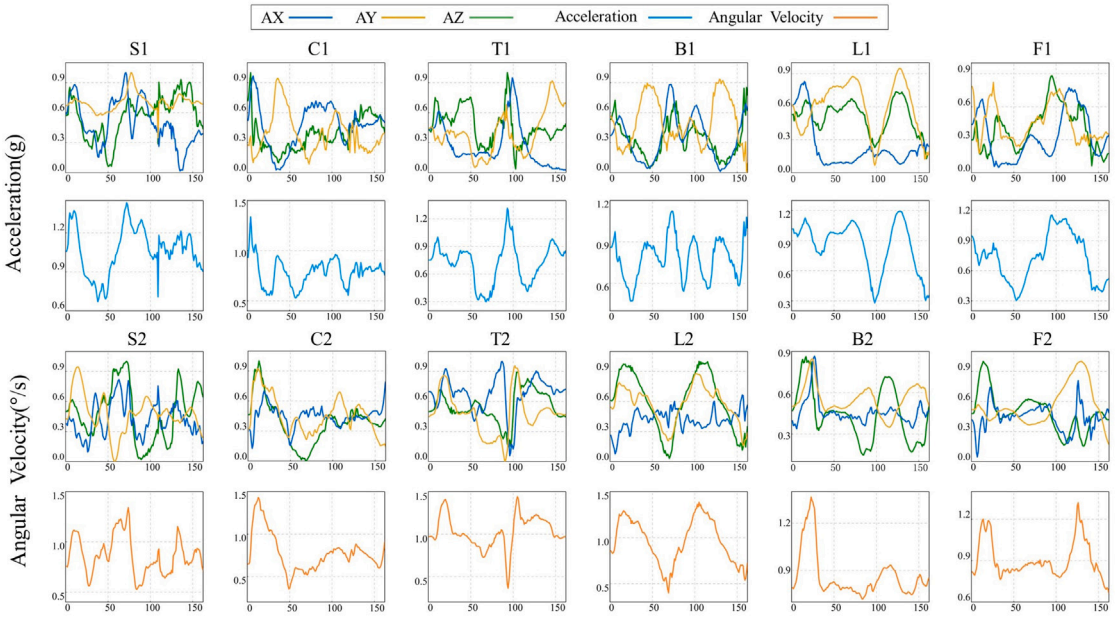

Fig. 2. The acceleration and angular velocity data of the 6 gymnastics movements measured at the left wrist point and the corresponding combined acceleration data and combined angular velocity data. AX, AY, and AZ represent data on the x-axis, y-axis, and z-axis, respectively.

be mitigated using data preprocessing techniques. As gymnastic action artifacts typically exhibit high-frequency characteristics, we chose a first-order low-pass filter, with the formula as follows:

$$Y(n) = \alpha X(n) + (1-\alpha)Y(n-1) \tag{3}$$

where α is the filter coefficient, $X(n)$ is the actual value of the nth data; $Y(n-1)$ is the output value of the previous filter; $Y(n)$ is the output value of the current filter.

We explored the values of α from 0 to 1 to find the optimal estimate of the data. When α is set to 0.4 (cut-off frequency = 3.18 Hz), the filter achieves the best results.

The experimental data of the left wrist position was chosen for the ultimate experiment. The data sets were bifurcated into an acceleration data set and an angular velocity data set, each comprising 1680 data sets. In turn, each data set was further divided into a training set and a test set, the ratio of which was 7:3. The final processed data is shown in Fig. 2.

2.3. Image encoding method

Converting 1D sensor data to images has three advantages. Firstly, the use of visual methods allows for a more expressive representation of the essence of the raw data and better expression of features related to HAR. Secondly, converting the raw data into three types of images – GAF, RP, and MTF – enables the extraction of features from the temporal, phase space, and state levels. Lastly, deep learning techniques have advantages in capturing and analyzing image features. Visualizing the data allows deep learning to better understand the intrinsic structure of the data and thus improve the accuracy of action recognition.

2.3.1. Gramian Angular Field (GAF)

The acceleration dataset and angular velocity dataset of human gymnastic actions are transformed into Gramian Angle Field (GAF) images. If $S = \{s_1, s_2, \ldots, s_i, s_j \ldots s_n\}$ is a time series data of a gymnastics movement with a data length of n, normalize S between 0 and 1 to get $\bar{S} = \{s'_1, s'_2 \ldots s'_i, s'_j \ldots s'_n\}$. Then it will be mapped to the angular coordinate system, and the mapping method is as follows:

$$\theta_i = arccos\left(s'_i\right) \tag{4}$$

where θ_i is the polar angle, which preserves the numerical information of the time series data, and s'_i represents the ith value of $\bar{S}$.

$$r_i = \frac{i}{n} \tag{5}$$

where r_i is the polar axis, which preserves the time information of the timing point data.

This encoding method enables the reconstruction of the initial time series based on angular coordinate system data. Furthermore, the correlation with the original time series data is preserved. Two types of images can be generated by GAF: Gram Sum Angle Field (GASF) and Gram Difference Angle Field (GADF). In this paper, the symmetric GASF is selected, and its formula is provided below:

$$\begin{bmatrix} \cos(\theta_1+\theta_1) & \cos(\theta_1+\theta_2) & \cdots & \cos(\theta_1+\theta_n) \\ \cos(\theta_2+\theta_1) & \cos(\theta_2+\theta_2) & \cdots & \cos(\theta_2+\theta_n) \\ \vdots & \vdots & \vdots & \\ \cos(\theta_n+\theta_1) & \cos(\theta_n+\theta_2) & \cdots & \cos(\theta_n+\theta_n) \end{bmatrix} = \left(S'\right)^T S' - \sqrt{E-(S')^2}^T \sqrt{E-(S')^2} \tag{6}$$

where E is the identity matrix.

The two datasets containing six gymnastic actions are transformed into GAF images, as depicted in Fig. 3.

2.3.2. Recurrence Plot (RP)

In order to assess the stationarity and internal similarity of the time series data, the acceleration dataset and angular velocity dataset were transformed into Recurrence Plot (RP) images. For $\bar{S}$ in 2.3.1, the formula for converting to RP is as follows:

$$RP = \alpha(\lambda - \| s(i) - s(j) \|) \tag{7}$$

where λ is the threshold and α is the coefficient. $s(i)$ represents the transformation of the ith point in the time domain into the corresponding state of the phase space and $s(j)$ represents the transformation of the jth point in the phase space.

The two datasets containing six gymnastic actions are transformed into RP images, as depicted in Fig. 3.

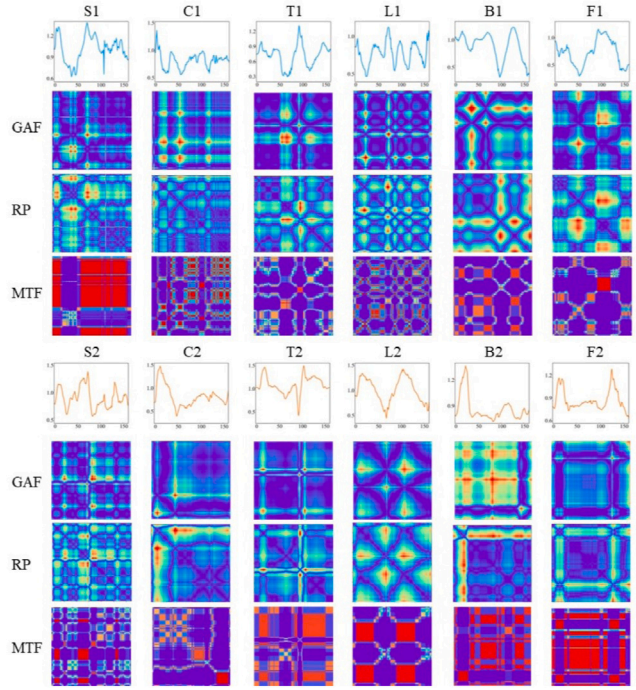

Fig. 3. GAF, RP and MTF images of Acceleration dataset and Angular velocity dataset.

2.3.3. Markov transition field (MTF)

For $\bar{S}$ in 2.3.1, firstly, The data quantiles are established based on the value range of $\bar{S}$, and the number of quantiles is B. Each data quantile belongs to a unique interval, which is referred to as b_i ($i \in [1, B]$). Then,a weighted adjacency matrix of size $[B, B]$ is constructed, denoted as W. To compensate for information loss, W is transformed into a Markov transition field (MTF) matrix of size $[n, n]$, denoted as M. The expression for matrix M is given as follows:

$$M = \begin{bmatrix} W_{ji|s'_1 \in b_j, s'_1 \in b_i} & W_{ji|s'_1 \in b_j, s'_2 \in b_i} & \cdots & W_{ji|s'_1 \in b_j, s'_n \in b_i} \\ W_{ji|s'_2 \in b_j, s'_1 \in b_i} & W_{ji|s'_2 \in b_j, s'_2 \in b_i} & \cdots & W_{ji|s'_2 \in b_j, s'_n \in b_i} \\ \vdots & \vdots & \vdots & \\ W_{ji|s'_n \in b_j, s'_1 \in b_i} & W_{ji|s'_n \in b_j, s'_2 \in b_i} & \cdots & W_{ji|s'_n \in b_j, s'_n \in b_i} \end{bmatrix} \tag{8}$$

where W_{ji} denotes the transfer frequency from quantile j to quantile i.

10 data quantiles are chosen to be employed due to the normalization of data within the range of 0 and 1. The two datasets containing six gymnastic actions were transformed into MTF images, as shown in Fig. 3.

When converting 1D time series into images using GAF, RP, and MTF methods, the entire sequence is used for conversion, and image resizing is subsequently performed using interpolation techniques.

2.4. ViTGS

At the outset of ViTGS, the 1D sensor data is converted into images. Subsequently, features are extracted from these images using the deep learning network VIT_B/16 and the efficient Gated Fusion Network (GFN) is used to fuse them together. Serving as the backbone of ViTGS, GFN is capable of effectively fusing these features. The fused features are then used as input by the SVM. The schematic of ViTGS is presented in Fig. 4a.

2.4.1. VIT_B/16

The features of GAF, RP, and MTF images are extracted by the ViT_B/16 deep learning network. The specific architecture of the network is depicted in Fig. 4b. ViT_B/16 performs exceptionally well in handling large amounts of data while saving computing resources. ViT_B/16 uses a convolution kernel size of 16 × 16 and a step size of 16. When an image with a size of [224, 224, 3] is inputted, the size of the image becomes [14, 14, 768] after convolution. Finally, it is flattened into a two-dimensional matrix with a size of [196, 768]. Prior to inputting this two-dimensional matrix into the Encoder, trainable parameters, including the vector Class Token and Position Embedding, are added for classification. After splicing the Class Token, the size of the two-dimensional matrix becomes [197,768]. Upon superimposing the Position Embedding, the matrix's size remains unchanged. Subsequently, the two-dimensional matrix is input into the Encoder Block.

The first Linear layer of the MLP Block in the Encoder Block expands the original input by four times, which is a crucial difference from the basic structure of the original Transformer's Encoder Block. For a 2D matrix of size [197,768], the output of this layer is [197.3072]. The second Linear layer reverts the output's size to [197,768]. The Encoder Block outputs a corresponding vector for each input vector. We only take the output vector of the Class Token corresponding to the Encoder Block, and the size is [1,768].

2.4.2. Gated Fusion Network (GFN)

The structure of the Gated Fusion Network (GFN) is depicted in Fig. 4c. The feature extracted from the Class Token of ViT_B/16 serves as the input to GFN. The features of GAF, RP, and MTF images, denoted as C1, C2, and C3, undergo convolution with the high-frequency lifting filter K and subsequently are processed by bilateral filtering.

The high-frequency lifting filter K is capable of precisely extracting crucial feature information while preserving low-frequency information and accentuating high-frequency information. The expression of K is as follows:

$$K = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & -1 \end{pmatrix} \tag{9}$$

Bilateral filtering is a nonlinear filtering method. The principle is to multiply a function related to the spatial distance by a function related to the gray distance. It can eliminate noise and enhance edges because it considers both spatial and grayscale information. The formula for bilateral filtering is as follows:

$$g(i,j) = \frac{\sum_{k,l} f(k,l)\, w(i,j,k,l)}{\sum_{k,l} w(i,j,k,l)} \tag{10}$$

$$w(i,j,k,l) = \exp\left(-\frac{(i-k)^2 + (j-l)^2}{2\sigma_d^2} - \frac{\| f(i,j) - f(k,l) \|^2}{2\sigma_r^2} \right) \tag{11}$$

where $f(i,j)$ are the coordinate of the current convoluted pixel, $f(k,l)$ are the coordinate of the neighboring pixels, $w(i,j,k,l)$ is the weighting coefficient, σ_d^2 and σ_r^2 is the standard deviation of the Gaussian function.

The last step in this process involves the normalization and activation of the extracted features through the Tanh function, resulting in gating weights W1, W2, and W3. The normalization formula utilized for this purpose is as follows:

$$y = \frac{x - mean[x]}{\sqrt{Var[x] + eps}} \times \gamma + \beta \tag{12}$$

where γ and β are the learning parameters, $Var[x]$ is the variance of x, and $mean[x]$ is the mean of x.

The formula for the activation function Tanh is presented as follows:

$$tanhx = \frac{sinhx}{coshx} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{13}$$

The fusion features can be obtained by serially multiplying the weights W1, W2, and W3 with the features C1, C2, and C3. These fusion features are then conveyed to the classifier (SVM) for classification.

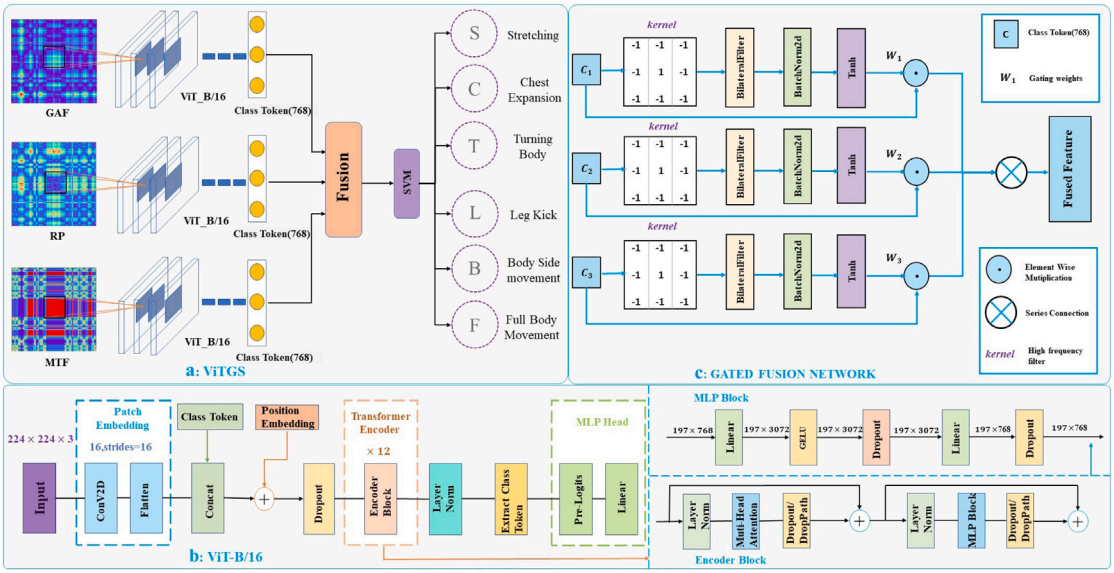

Fig. 4. The structure diagram of the three-channel convolution model ViTGS. Among them, a is the overall frame diagram of the model, b is the structure diagram of ViT_B/16, and c is the structure diagram of GFN.

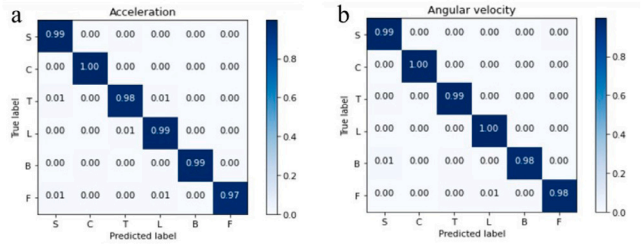

Fig. 5. The confusion matrix of the ViTGS classification results.

3. Experiment

In this study, our model's superiority was confirmed by conducting experiments on both the acceleration and angular velocity datasets. The programs were written in Python 3.6 using PyTorch and executed on a RTX 3080 GPU.

3.1. Experimental results of vitgs

The remarkable performance of ViTGS in analyzing 1D time series data has been demonstrated. The accuracy rates for classification of 6 gymnastic actions are 98.56% and 99.04% on the acceleration and angular velocity datasets, as indicated in Table 2. The confusion matrices of the ViTGS classification results are shown in Fig. 5a and b. It is evident that our model has a high classification accuracy rate for each action, with each action in the acceleration dataset exceeding 97%, and above 98% in the angular velocity dataset. Moreover, the recognition accuracy rate of chest expansion (C) is 100%, attesting to the model's ability to accurately recognize all 6 gymnastic actions.

3.2. Ablation experiment

The ViTGS has been refined from the 3-channel image fusion model. The model comprises three components: a deep learning network designed for image feature extraction; a feature fusion method which involves sequential features and a fully connected classifier. Our model has been enhanced in each parts and its superiority has been validated through experiments.

ViT_B/16: In order to determine the superiority of ViT_B/16 over other mainstream deep learning networks for feature extraction, experiments were conducted on the acceleration dataset using various networks such as ShuffleNet_v2_x0_5, MnasNet0_5, GooglNet, EfficientNet_b0, AlexNet, MobileNet_v2, and ResNet18. At the same time, Concatenated vectors were used as the feature fusion method, while a fully connected classifier was chosen. It was ensured that the only variable in the experiment was the deep learning network used. The outcomes of various networks are depicted in Fig. 6a and b. Fig. 6a presents the radar map of the evaluation outcomes on the test dataset with different deep learning networks. It is apparent that the highest values in the classification evaluation metrics, including Accuracies,

GFN: Once ViT_B/16 was determined as the deep learning network, ViTGS used GFN to fuse features. Compared to the method of fusing features by concatenating one-dimensional vectors, GFN enhances the differences and preserves the integrity of features, achieving better performance. The experiment results are displayed in Table 2. By replacing the concatenated vector with GFN and unifying the classifier to fully connected, the improved model achieved a classification accuracy of 98.07% on the test set, which represents a 0.56% increase compared to 97.52% before the improvement. The other evaluation indicators, including Precision, Recall, and F1-score, also exhibited improvements, with values exceeding 98%. Thus, the proposed feature fusion framework GFN was found to be more effective.

SVM: After comparing various classifiers, including fully connected (FC), K-Nearest Neighbor (KNN), RandomForest, and SVM, the best-performing SVM was selected as the classifier, as shown in the detailed results in Table 2. The SVM classifier achieved the highest accuracy rate of 98.56%, which was 0.50% higher than FC's accuracy of 98.07%. Additionally, other evaluation metrics such as precision, recall, and F1-score were also the highest for SVM. Therefore, SVM was chosen as the classifier.performance on the acceleration dataset. Similarly, the performance on the angular velocity dataset was evaluated using the same process as mentioned earlier. The outcomes of the different schemes on the training and test sets are shown in a and c of Fig. 6. Ultimately, Table 2 indicates that the highest accuracy rate of 99.04% was achieved by the ViTGS, with the rest of the indicators also being the highest.

3.3. Comparison with different data preprocessing methods

We attempted three different data preprocessing approaches. The first approach involved subjectively segmenting the data based on

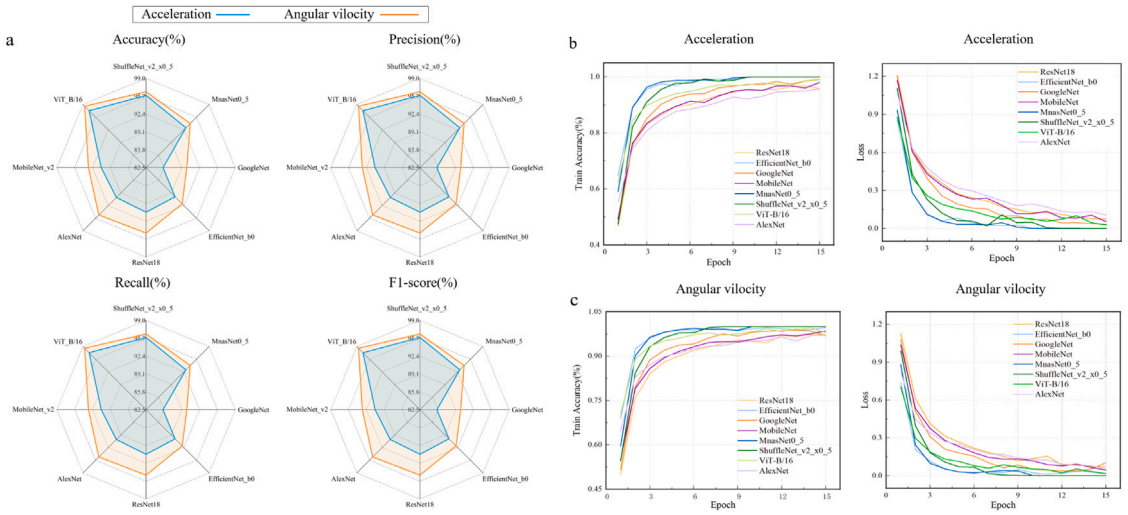

Fig. 6. Classification result of different networks (a) and loss and accuracy curves on the acceleration dataset (b) and angular velocity dataset (c).

Table 2
The experimental results of the data set in ablation experiment.

Dataset	Net	Fusion	Classification	Accuracy	Precision	Recall	F1-score
Acceleration	ViT_B /16	Concatenate	FC	0.9752	0.9752	0.9752	0.9750
		GFN	FC	0.9807	0.9805	0.9805	0.9803
		GFN	RandomForest	0.9289	0.9289	0.9287	0.9289
		GFN	KNN	0.9507	0.9503	0.9507	0.9507
		GFN	SVM	0.9856	0.9855	0.9856	0.9855
Angular velocity	ViT_B /16	Concatenate	FC	0.9848	0.9850	0.9848	0.9848
		GFN	FC	0.9863	0.9863	0.9863	0.9861
		GFN	RandomForest	0.9463	0.9461	0.9463	0.9463
		GFN	KNN	0.9715	0.9712	0.9715	0.9715
		GFN	SVM	0.9904	0.9903	0.9904	0.9903

domain experts' experience, followed by feature extraction and machine learning classification. The second approach involved directly inputting the merged 6-axis IMU data into a 1D-CNN for automatic feature extraction and classification using deep learning. The third approach, which is the novelty of this paper, involved transforming the one-dimensional time series data into three types of images (GAF, RP, MTF) using the image-based method. Subsequently, our proposed ViTGS model was used for feature extraction, fusion, and classification.

Specifically, for the first method, we manually segmented the averaged values of the acceleration and angular velocity data for each action into 60 groups, obtaining the standard waveforms for each action. Then, we extracted motion features from the waveforms representing the standard actions, including mean, standard deviation, mean square error, information entropy, maximum value, minimum value, and area. Among them, the maximum and minimum values reflect the intensity of gymnastics exercises, the mean value reflects the trend of motion data during gymnastics exercises, the standard deviation reflects the dispersion of gymnastics motion data during execution, and the mean square error is used to evaluate the accuracy and stability of actions. Information entropy reflects the diversity and complexity of actions, while area reflects the continuity and changes in actions. Finally, we employed four machine learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART) to identify six types of gymnastics actions. To avoid overfitting, we adopted cross-validation.

Regarding the second method, we conducted experiments using the classic one-dimensional time series classification approach. We directly input the accelerometer dataset and gyroscope dataset into a one-dimensional convolutional neural network for training and classification. The deep learning networks used in this experiment include AlexNet, DenseNet, GoogleNet, LeNet, ResNet50, VGG19, and ZFNet.

Table 3
Classification results of three different data preprocessing methods.

Method	Model	Accuracy (Acc)	Accuracy (Gyro)
Expert segmentation, machine learning	KNN	95.0%	96.6%
	NB	90.6%	92.5%
	LDA	98.1%	96.8%
	CART	88.8%	89.4%
1D-CNN	AlexNet	93.7%	95.8%
	DenseNet	89.3%	95.8%
	GoogleNet	87.0%	82.2%
	LeNet	82.8%	90.3%
	ResNet50	90.9%	97.6%
	VGG19	96.8%	97.6%
	ZFNet	95.3%	97.2%
Our work	ViTGS	98.6%	99.0%

For comparison, we used motion data from the left wrist for recognition. We designated 80% of the dataset as the training set and the remaining 20% as the test set. The results shown in Table 3 indicate that using different data preprocessing methods will lead to changes in accuracy, and the data preprocessing methods used in our work have the best accuracy.

3.4. Comparing with other work

In this section, we have conducted a comprehensive comparative analysis of our work with other researchers in the field of sports action recognition. The result is shown in Table 4. Our goal is to evaluate the effectiveness of our method in the context of existing approaches.

Compared to action recognition techniques based on video data and single-node sensors, our method demonstrates clear advantages in classification accuracy. This indicates that our method is able to recognize

Table 4
Previous research work related to video and sensor-based sports action recognition.

Type	Literature	Sensor type	Node	Position	Number of activities	Method	Accuracy
Video	Ji (2020)	–	–	–	4 basketball activities (e.g., shooting, passing, catching, dribbling)	Image feature extraction and Bayesian algorithm	91.28%
	Peng and Zhang (2021)	–	–	–	11 dancing actions (e.g., Left arm raised, arms downward, clap, etc.)	Generating skeleton node graph	93.43%
	Qiao and Shen (2021)	–	–	–	5 dancing actions (e.g., belly dance, Latin dance, ballet dance, national dance, folk dance)	Hybrid model of human motion feature	90.04%
Single-sensor node	Xia et al. (2020)	Acc, Gyro	1	Wrist	5 badminton activities (e.g., badminton stroke, badminton smash, etc.)	K-means clustering, density-based spatial clustering	94.69%
	Shahar et al. (2020)	Acc, Gyro	1	Waist	6 hockey activities (e.g., passing, drive, drag, etc.)	Best window segmentation size, SVM	91.40%
	Wang et al. (0000)	Acc, Gyro	1	Waist	4 competitive swimming styles (e.g., butterfly, breaststroke, freestyle, back crawl)	Hidden Markov Model (HMM)	94.38%
	Delhaye et al. (2022)	Acc, Gyro	1	Sacrum	8 swimming actions (e.g., wall push, butterfly, turn, backstroke, front crawl)	Sliding window overlapped frame, DCNN network	96.00%
	Zhao et al. (2021)	Acc, Gyro	1	Left hand	6 gymnastic movements (e.g., stretching, chest expansion, twisting, etc.)	Linear discriminant analysis (LDA)	97.31%
Multiple-sensor node	Worsey et al. (2020)	Acc, Gyro	2	Waists	5 boxing activities (e.g., left jab, right cross, etc.)	SVM with a Gaussian rbf kernel	96.00%
	Ni and Yao (2021)	Acc, Gyro	3	Body	4 dancing activities (e.g., linear movement, rotating, etc.)	SVM	92.00%
	Pajak et al.	Acc, Gyro	3	Chest, hand and foot	3 sports activities (e.g., pull-ups, squats, dips)	CNN network	94.70%
	Jiao et al. (2018)	Acc, Gyro	2	Golf club	9 golf swing actions (e.g., straight, hook, pull, etc.)	DCNN network	95.05%
Single-sensor node	Our Work	Acc/Gyro	1	Left wrist	6 gymnastic movements(e.g., stretching, chest expansion, etc.)	Image encoding, ViTGS	98.56%/99.04%

and classify sports actions more accurately. This can be attributed to our imagery approach and innovative feature extraction and fusion techniques, which assist us in better extracting and utilizing information from sensor data that is beneficial for human activity recognition (HAR). Compared to multi-node sensor systems, our method performs exceptionally well, surpassing most of the previous studies. It is worth noting that we utilize a relatively small number of sensor nodes. This difference gives us a dual advantage: simplified operational complexity and higher action recognition rate. These factors significantly contribute to the feasibility and practicality of deploying our method in real-world applications.

4. Conclusion

HAR based on wearable sensors holds significant value in applications such as gymnastics action recognition. Compared to video-based methods, sensor-based approaches offer advantages such as low-cost experimentation, unconstrained data collection, and protection of participants' privacy. However, the feature extraction and fusion of sensor data remain challenging tasks. This paper aims to address the issues related to feature extraction and fusion of sensor data by proposing an innovative action recognition framework that combines time-series data imaging techniques with a three-channel convolutional model called ViTGS. By adopting three imaging methods, namely GAF, RP, and MTF, we can capture the features of time-series data from multiple perspectives and generate three-channel images. The ViTGS model utilizes the ViT_B/16 network to extract image features and effectively fuse different channel features using GFN. By employing SVM for classification, we achieve high accuracy in gymnastics action recognition. Experiments were conducted using a gymnastics action dataset we collected. The results demonstrate that in the left-wrist single-node experiment, we achieved an accuracy of 98.56% on the acceleration dataset, which is 2.8% higher than the optimal accuracy (95.70%). Additionally, an accuracy of 99.04% was obtained on the angular velocity dataset, which is 2.50% higher than the optimal accuracy (96.62%). The findings of this study hold significant implications for practical applications. Our method can assist gymnastics coaches and athletes in real-time action evaluation and feedback, thereby enhancing training effectiveness and competitive performance.

Furthermore, our research findings have practical applications in multiple fields. For example, in the field of sports training, it can achieve personalized customization, providing athletes with more precise guidance and feedback. In terms of medical rehabilitation monitoring, it can track the progress of rehabilitation in real-time and provide personalized rehabilitation advice for patients. It can also improve production efficiency in industrial production optimization and be applied in the field of virtual reality. However, our method also has some limitations. The complexity of the ViTGS model may require more time and effort during deployment or reduce its real-time performance in practical applications. Additionally, the model has relatively high hardware resource requirements, which may be limited by hardware constraints in certain situations, restricting its application scope in specific environments. In the future, we will continue to work on improving our model, striving to reduce its complexity to enhance deployment efficiency and real-time performance. Through continuous improvement and innovation, we hope to better meet the needs of practical applications and bring greater value and potential to various fields.

CRediT authorship contribution statement

Yuliang Zhao: Conceptualization, Formal analysis, Investigation, Resources, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Fanghecong Dong:** Conceptualization, Methodology, Validation, Software, Investigation, Data curation, Writing – original draft. **Tianang Sun:** Investigation, Resources. **Zhongjie Ju:** Investigation, Resources. **Le Yang:** Investigation, Resources. **Peng Shan:** Methodology, Validation. **Lianjiang Li:** Investigation, Resources. **Xiaoyong Lv:** Formal analysis, Resources, Writing – review & editing. **Chao Lian:** Conceptualization, Formal analysis, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61873307), the Hebei Natural Science Foundation (Grant Nos. F2020501040, F2021203070, and F2021501021), the Fundamental Research Funds for the Central Universities under Grant N2123004, the Administration of Central Funds Guiding the Local Science and Technology Development (Grant No. 206Z1702G).

References

Ahmed, N., Rafiq, J. I., & Islam, M. R. (2020). Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model. *Sensors*, *20*(1), 317.

Alawneh, L., Alsarhan, T., Al-Zinati, M., Al-Ayyoub, M., Jararweh, Y., & Lu, H. (2021). Enhancing human activity recognition using deep learning and time series augmented data. *Journal of Ambient Intelligence and Humanized Computing*, 1–16.

Bhattacharya, D., Sharma, D., Kim, W., Ijaz, M. F., & Singh, P. K. (2022). Ensem-HAR: An ensemble deep learning model for smartphone sensor-based human activity recognition for measurement of elderly health monitoring. *Biosensors*, *12*(6), 393.

Challa, S. K., Kumar, A., & Semwal, V. B. (2022). A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. *The Visual Computer*, *38*(12), 4095–4109.

Delhaye, E., Bouvet, A., Nicolas, G., Vilas-Boas, J. P., Bideau, B., & Bideau, N. (2022). Automatic swimming activity recognition and lap time assessment based on a single IMU: a deep learning approach. *Sensors*, *22*(15), 5786.

Dhiman, C., & Vishwakarma, D. K. (2019). A robust framework for abnormal human action recognition using $\mathcal{R}$-Transform and Zernike moments in depth videos. *IEEE Sensors Journal*, *19*(13), 5195–5203.

Dua, N., Singh, S. N., & Semwal, V. B. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. *Computing*, *103*(7), 1461–1478.

Gholamiangonabadi, D., Kiselov, N., & Grolinger, K. (2020). Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *IEEE Access*, *8*, 133982–133994.

Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights*, *1*(2), Article 100046.

Haider, F., Salim, F. A., Postma, D. B., Van Delden, R., Reidsma, D., van Beijnum, B.-J., & Luz, S. (2020). A super-bagging method for volleyball action recognition using wearable sensors. *Multimodal Technologies and Interaction*, *4*(2), 33.

Han, C., Zhang, L., Tang, Y., Huang, W., Min, F., & He, J. (2022). Human activity recognition using wearable sensors by heterogeneous convolutional neural networks. *Expert Systems with Applications*, *198*, Article 116764.

Huang, G., & Yan, Q. (2020). Optimizing features quality: a normalized covariance fusion framework for skeleton action recognition. *IEEE Access*, *8*, 211869–211881.

Huang, W., Zhang, L., Gao, W., Min, F., & He, J. (2021). Shallow convolutional neural networks for human activity recognition using wearable sensors. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Jalal, A., Uddin, M. Z., & Kim, T.-S. (2012). Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Transactions on Consumer Electronics*, *58*(3), 863–871.

Jameer, S., & Syed, H. (2023). Deep SE-BiLSTM with IFPOA fine-tuning for human activity recognition using mobile and wearable sensors. *Sensors*, *23*(9), 4319.

Ji, R. (2020). Research on basketball shooting action based on image feature extraction and machine learning. *IEEE Access*, *8*, 138743–138751.

Jiao, L., Wu, H., Bie, R., Umek, A., & Kos, A. (2018). Multi-sensor golf swing classification using deep CNN. *Procedia Computer Science*, *129*, 59–65.

Lee, J., & Ahn, B. (2020). Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors*, *20*(10), 2886.

Lian, C., Ma, R., Wang, X., Zhao, Y., Peng, H., Yang, T., Zhang, M., Zhang, W., Sha, X., & Li, W. J. (2021). ANN-enhanced IoT wristband for recognition of player identity and shot types based on basketball shooting motion analysis. *IEEE Sensors Journal*, *22*(2), 1404–1413.

Lian, C., Ren, X., Zhao, Y., Zhang, X., Chen, R., Wang, S., Sha, X., & Li, W. J. (2020). Towards a virtual keyboard scheme based on wearing one motion sensor ring on each hand. *IEEE Sensors Journal*, *21*(3), 3379–3387.

Lu, Y., Wei, Y., Liu, L., Zhong, J., Sun, L., & Liu, Y. (2017). Towards unsupervised physical activity recognition using smartphone accelerometers. *Multimedia Tools and Applications*, *76*(8), 10701–10719.

Mekruksavanich, S., & Jitpattanakul, A. (2021). Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. *Electronics*, *10*(3), 308.

Mim, T. R., Amatullah, M., Afreen, S., Yousuf, M. A., Uddin, S., Alyami, S. A., Hasan, K. F., & Moni, M. A. (2023). GRU-INC: An inception-attention based approach using GRU for human activity recognition. *Expert Systems with Applications*, *216*, Article 119419.

Morales, J., & Akopian, D. (2017). Physical activity recognition by smartphones, a survey. *Biocybernetics and Biomedical Engineering*, *37*(3), 388–400.

Nafea, O., Abdul, W., Muhammad, G., & Alsulaiman, M. (2021). Sensor-based human activity recognition with spatio-temporal deep learning. *Sensors*, *21*(6), 2141.

Ni, S., & Yao, D. (2021). Sports dance action recognition system oriented to human motion monitoring and sensing. *Wireless Communications and Mobile Computing*, *2021*, 1–10.

Niemelä, M., Kiviniemi, A., Ikäheimo, T. M., Tulppo, M., Korpelainen, R., Jämsä, T., & Farrahi, V. (2023). Compositional association of 24 h movement behavior with incident major adverse cardiac events and all-cause mortality. *Scandinavian Journal of Medicine & Science in Sports*, *33*(5), 641–650.

Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*, 147–170.

Pajak, I., Krutz, P., Patalas-Maliszewska, J., Rehm, M., Pajak, G., Schlegel, H., & Dix, M. Sports activity recognition with UWB and inertial sensors using deep learning approach. In *2022 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–8). IEEE.

Peng, F., & Zhang, H. (2021). Research on action recognition method of dance video image based on human-computer interaction. *Scientific Programming*, *2021*, 1–9.

Qiao, L., & Shen, Q. (2021). Human action recognition technology in dance video image. *Scientific Programming*, *2021*, 1–8.

Qin, Z., Zhang, Y., Meng, S., Qin, Z., & Choo, K.-K. R. (2020). Imaging and fusing time series for wearable sensor-based human activity recognition. *Information Fusion*, *53*, 80–87.

Ramanujam, E., Perumal, T., & Padmavathi, S. (2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, *21*(12), 13029–13040.

Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235–244.

Sarkar, A., Hossain, S. S., & Sarkar, R. (2023). Human activity recognition from sensor data using spatial attention-aided CNN with genetic algorithm. *Neural Computing and Applications*, *35*(7), 5165–5191.

Sena, J., Barreto, J., Caetano, C., Cramer, G., & Schwartz, W. R. (2021). Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble. *Neurocomputing*, *444*, 226–243.

Sha, X., Lian, C., Zhao, Y., Yu, J., & Li, W. J. (2020). An explicable keystroke recognition algorithm for customizable ring-type keyboards. *IEEE Access*, *8*, 22933–22944.

Shahar, N., Ghazali, N. F., As ari, M. A., Tan, T. S., & Ibrahim, M. F. (2020). Investigation of different time-series segmented windows from inertial sensor for field hockey activity recognition. In *Enhancing health and sports performance by design: proceedings of the 2019 movement, health & exercise (MoHE) and international sports science conference (ISSC)* (pp. 299–310). Springer.

Shayesteh, M. H., Shahrokhzadeh, B., & Masoumi, B. (2023). Game theory solutions in sensor-based human activity recognition: A review. *Journal of AI and Data Mining*, *11*(2), 259–289.

Shi, L.-F., Liu, Z.-Y., Zhou, K.-J., Shi, Y., & Jing, X. (2023). Novel deep learning network for gait recognition using multimodal inertial sensors. *Sensors*, *23*(2), 849.

Siddiqi, M. H., Almashfi, N., Ali, A., Alruwaili, M., Alhwaiti, Y., Alanazi, S., & Kamruzzaman, M. (2021). A unified approach for patient activity recognition in healthcare using depth camera. *IEEE Access*, *9*, 92300–92317.

Sun, P. S., Xu, D., Mai, J., Zhou, Z., Agrawal, S., & Wang, Q. (2020). Inertial sensors-based torso motion mode recognition for an active postural support brace. *Advanced Robotics*, *34*(1), 57–67.

Thirunavukarasu, R., Gnanasambandan, R., Gopikrishnan, M., & Palanisamy, V. (2022). Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. *Computers in Biology and Medicine*, Article 106020.

Thu, N. T. H., & Han, D. S. (2021). HiHAR: A hierarchical hybrid deep learning architecture for wearable sensor-based human activity recognition. *IEEE Access*, *9*, 145271–145281.

Wang, Y., Cang, S., & Yu, H. (2019). A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, *137*, 167–190.

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11.

Wang, H., & Hu, S. (2021). Gymnastics movement signs based on network communication and body contour feature extraction. *Mobile Information Systems*, *2021*, Article 8336367. http://dx.doi.org/10.1155/2021/8336367.

Wang, L., & Liu, R. (2020). Human activity recognition based on wearable sensor using hierarchical deep LSTM networks. *Circuits, Systems, and Signal Processing*, *39*(2), 837–856.

Wang, Z., Shi, X., Wang, J., Gao, F., Li, J., Guo, M., Zhao, H., & Qiu, S. Swimming motion analysis and posture recognition based on wearable inertial sensors. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*, IEEE (pp. 3371–3376).

Worsey, M. T., Espinosa, H. G., Shepherd, J. B., & Thiel, D. V. (2020). An evaluation of wearable inertial sensor configuration and supervised machine learning models for automatic punch classification in boxing. *IoT*, *1*(2), 360–381.

Xia, K., Wang, H., Xu, M., Li, Z., He, S., & Tang, Y. (2020). Racquet sports recognition using a hybrid clustering model learned from integrated wearable sensor. *Sensors*, *20*(6), 1638.

Xu, J., & Yuan, K. (2021). Wearable muscle movement information measuring device based on acceleration sensor. *Measurement*, *167*, Article 108274.

Yadav, S. K., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, *223*, Article 106970.

Yen, C.-T., Liao, J.-X., & Huang, Y.-K. (2020). Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms. *Ieee Access*, *8*, 174105–174114.

Zaitseva, E., Levashenko, V., Rabcan, J., & Kvassay, M. (2023). A new fuzzy-based classification method for use in smart/precision medicine. *Bioengineering*, *10*(7), 838.

Zhang, M., Li, H., Ge, T., Meng, Z., Gao, N., & Zhang, Z. (2022). Integrated sensing and computing for wearable human activity recognition with MEMS IMU and BLE network. *Measurement Science Review*, *22*(4), 193–201.

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, *22*(4), 1476.

Zhang, Z., Wang, W., An, A., Qin, Y., & Yang, F. (2023). A human activity recognition method using wearable sensors based on convtransformer model. *Evolving Systems*, 1–17.

Zhang, G., Wang, Z., Zhao, L., Qi, Y., & Wang, J. (2017). Coal-rock recognition in top coal caving using bimodal deep learning and Hilbert-Huang transform. *Shock and Vibration*, *2017*.

Zhao, Y., Lian, C., Zhang, X., Sha, X., Shi, G., & Li, W. J. (2019). Wireless IoT motion-recognition rings and a paper keyboard. *IEEE Access*, *7*, 44514–44524.

Zhao, Y., Meng, X., Ren, X., Lian, C., Sun, J., Sha, X., Wang, S., Li, L., & Li, W. J. (2021). Quantitative evaluation of gymnastics based on multiple MEMS sensors. *IEEE Sensors Journal*, *21*(21), 24531–24539.