BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 13/15 pp188–204
DOI: 10.26599/BDMA.2023.9020018
Volume 7, Number 1, March 2024

Attention-Based CNN Fusion Model for Emotion Recognition During Walking Using Discrete Wavelet Transform on EEG and Inertial Signals

Yan Zhao, Ming Guo*, Xiangyong Chen, Jianqiang Sun, and Jianlong Qiu

Abstract: Walking as a unique biometric tool conveys important information for emotion recognition. Individuals in different emotional states exhibit distinct walking patterns. For this purpose, this paper proposes a novel approach to recognizing emotion during walking using electroencephalogram (EEG) and inertial signals. Accurate recognition of emotion is achieved by training in an end-to-end deep learning fashion and taking into account multi-modal fusion. Subjects wear virtual reality head-mounted display (VR-HMD) equipment to immerse in strong emotions during walking. VR environment shows excellent imitation and experience ability, which plays an important role in awakening and changing emotions. In addition, the multi-modal signals acquired from EEG and inertial sensors are separately represented as virtual emotion images by discrete wavelet transform (DWT). These serve as input to the attention-based convolutional neural network (CNN) fusion model. The designed network structure is simple and lightweight while integrating the channel attention mechanism to extract and enhance features. To effectively improve the performance of the recognition system, the proposed decision fusion algorithm combines Critic method and majority voting strategy to determine the weight values that affect the final decision results. An investigation is made on the effect of diverse mother wavelet types and wavelet decomposition levels on model performance which indicates that the 2.2-order reverse biorthogonal (rbio2.2) wavelet with two-level decomposition has the best recognition performance. Comparative experiment results show that the proposed method outperforms other existing state-of-the-art works with an accuracy of 98.73%.

Key words: walking; multi-modal fusion; virtual reality; emotion recognition; discrete wavelet transform; attention mechanism

1 Introduction

The development of artificial intelligence is accompanied by technological and economic advances, with human-computer interaction becoming more frequent. Since most of the information exchanged and disseminated is emotional, people expect the computer to understand and process emotional information in addition to simply operating with the mouse and keyboard, which can lead to positive effects. Emotion recognition technology is to enable the computer to realize the ability to receive signals representing human emotions and infer emotional states to achieve human-centered human-computer interaction. As the key part

• Yan Zhao is with the School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the School of Automation and Electrical Engineering, Linyi University, Linyi 276000, China. E-mail: 1173243632@163.com.

• Ming Guo, Xiangyong Chen, Jianqiang Sun, and Jianlong Qiu are with the School of Automation and Electrical Engineering, Linyi University, Linyi 276000, China. E-mail: guoming0537@126.com; cxy8305@163.com; sjqyjs@sina.com; qiujianlong@lyu.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2022-11-03; revised: 2023-06-09; accepted: 2023-06-25

of affective computing, emotion recognition plays an important role in health care[1], educational application[2], and social research[3].

Previous research on emotion recognition based on facial expressions and speech signals has been already mature, but this is considered challenging mainly because the source may not be reliable in some cases. People can cover up real emotions by actively controlling expressions and speech. For walking, it is considered to be a promising source of emotional information with natural characteristics of non-contact and hard-to-camouflage, and there is evidence that human emotions are expressed to some extent through walking[4]. The posture and movement pattern during walking have obvious individual differences and uniqueness. Additionally, walking-based research avoids many of the privacy concerns associated with facial and speech emotion recognition systems. The research on walking motion recognition is the premise of emotion recognition during walking, that has applications in pedestrian navigation[5, 6], distance estimation[7], and criminal investigation monitoring[8]. The change of emotion leads to the change of walking pattern, and the mapping relationship between the two provides a reliable source for automatic emotion recognition. For example, partners tend to walk quickly when they reunite after a long absence, which is a manifestation of positive emotions. People usually have such walking pattern when they encounter happy or desirable things. On the contrary, people are prone to tension in terrible scenes so that they walk slowly, and even the body becomes uncontrolled and stationary if the emotions persist and intensify.

Access to emotional information generally relies on visual-based methods and wearable sensors. Capturing motion information is intuitive with related devices such as depth cameras, but in some cases, the monitoring range and shooting quality are easily limited. The development of wearable sensors provides a new non-invasive solution for emotion recognition[9]. The sensor can be flexibly installed in some parts of the human body, and also addresses the privacy challenges faced by long-term exposed to the camera. Wearable sensors for emotion recognition mainly includes physiological sensors and inertial sensors. The former can capture weak physiological signals controlled by the nervous system, which is reliable and objective. The latter, such as inertial measurement units (IMUs), can detect and measure acceleration, angular velocity, and multiple degrees of freedom (DOF) motion. It is worth noting that emotional expression is a diversified process, and research on single modality is not enough. With the improvement of multi-source heterogeneous information fusion theory, the complementation and promotion of multi-modality information can make up for the defects of single modality[10]. Therefore, multi-modality emotion recognition based on multi-type sensor fusion has gradually become a research hotspot[11].

The premise of realizing emotion recognition during walking is to define emotion, and the emotion model describes the expressive characteristics of emotional state. Ekman et al.[12] found that volunteers in five different cultural backgrounds showed extremely similar reactions in the six basic emotions of anger, disgust, fear, happiness, sadness, and surprise. The discrete emotion model can represent emotion intuitively, but it is hard to satisfy the complexity of emotion. The dimensional emotion model describes the specific continuity of emotion, which can represent the intensity and change process. Russell[13] proposed a two-dimensional emotion model that can represent the vast majority of emotions using valence and arousal. By adding the dominance dimension, the pleasure-arousal-dominance (PAD) model[14] with high acceptance was proposed, which can theoretically rexpress infinite emotions.

In recent years, the use of machine learning algorithms to solve the problem of emotion recognition has received extensive attention. The traditional recognition algorithm[15, 16] is necessary to manually extract and construct feature matrix as the input of the classifier. These features are generally empirical and difficult to fully reflect the original data. With the rapid development of deep learning, deep neural networks (DNN) has shown superior performance and can learn valuable features directly from input data without manual extraction. Currently, deep learning techniques have gained extensive research in emotion recognition. Aslan[17] converted electroencephalogram (EEG) signals into EEG images including time-frequency domain information, and used pre-trained GoogLeNet to extract features from EEG images. Atanassov et al.[18] used the pre-training DNN model to extract emotion from body posture and use specific datasets for training, which expanded on their previous research on facial emotion recognition. Yin et al.[19] proposed a new deep learning based emotion recognition method using EEG sensors. They segmented the calibrated EEG data, extracted the differential entropy features, and constructed the feature cube as the input of the graph convolutional network (GCN) and the long short-term memory (LSTM) neural network. Fan et al.[20] proposed a deep convolutional neural network with attention mechanism for electrocardiogram (ECG) emotion recognition, in which the attention mechanism learned weights from the ECG features extracted by convolutional neural network (CNN).

The performance of CNN is confirmed by the automatic extraction of deep features, and multi-modal fusion further achieves accurate real-time recognition of emotions in real-world scenarios. So this paper proposes an attention-based CNN fusion model for emotion recognition during walking using discrete wavelet transform on EEG and inertial signals. Figure 1 provides an overview of all methods used and the entire process. First, the subjects wearing the virtual reality head-mounted display (VR-HMD) equipment are guided to generate an immersive feeling to stimulate emotions, while EEG and inertial sensors begin to collect walking-motion data. Then, the multi-modal data are separately processed and represented as images by discrete wavelet transform (DWT) as input to the attention-based CNN fusion model to extract features and realize emotion recognition during walking.

The main contributions and innovations of this paper are as follows:

- Overall, we propose a new approach to recognizing emotion during walking using inertial and EEG signals, which takes into account multi-modal fusion and trains in an end-to-end deep learning fashion to achieve accurate emotion recognition.
- Regarding to the sensor data representation, a simple and effective feature transformation method based on DWT is designed to represent the input signals as time-frequency domain.
- Regarding to the deep feature extraction, we developed a CNN structure combined with channel attention mechanism. CNN extracts discriminative features according to the correlation of multi-channel signals in the sensors. The channel attention mechanism adaptively emphasizes the key parts inside the feature map to achieve further feature optimization.
- In order to achieve multi-modal fusion, a decision fusion algorithm is proposed, which uses the evaluation matrix and Critic method to assign weights to the prediction labels that may affect the final decision. According to the majority voting strategy, the final prediction result of emotion during walking is obtained.
- In order to stimulate real and profound emotions, this paper uses VR-HMD equipment to enable subjects to immersely interact with the virtual environment to generate happy and fear emotion.

2 Material and Method

2.1 Data collection

This paper uses EEG sensors and inertial sensors to perceive emotions from internal and external manifestations. The former records electrophysiological indicators of brain activity, while the latter captures walking motion to obtain close-range data. According to the relevant literature[21, 22], we install two inertial sensors and an EEG sensor on the thighs and head of each subject respectively after comprehensive consideration, which can realistically

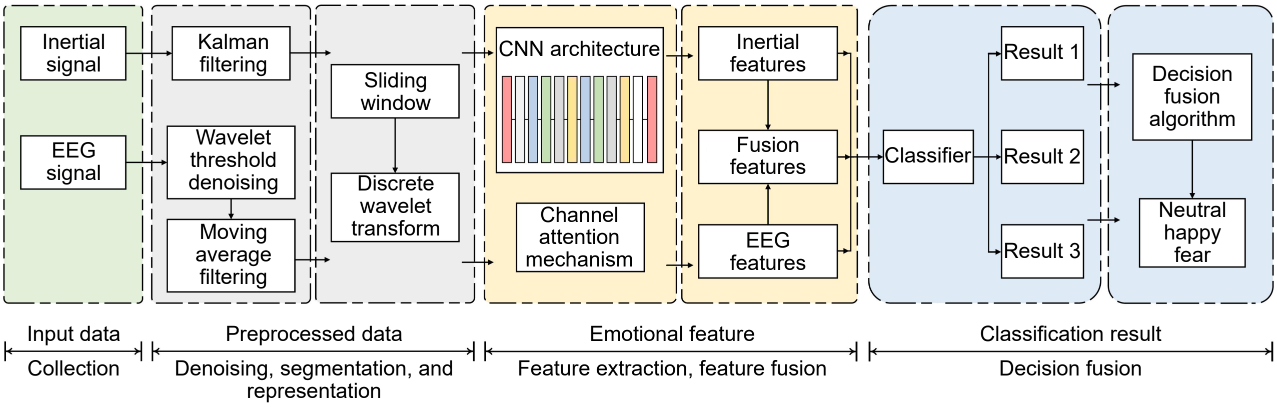


Fig. 1 Overview of all the methods and the entire process used in emotion recognition during walking.

simulate the real scene of people putting their smartphones in their trouser pockets. Each inertial sensor (WIT Inc., CHN) has built-in ICM42605 (3-axis accelerometer and 3-axis gyroscope) and MMC3630 (3-axis magnetometer), which have the characteristics of small size, wearable, and low power consumption[23, 24]. The specific parameters of inertial sensor are shown in Table 1. The EEG sensor (SICHIRAY Inc., CHN) has built-in ThinkGear Asic Module (TGAM) and 50 Hz notch filter, which adopts serial communication protocol and outputs EEG data at 9600 baud rate. The sampling frequencies of inertial sensor and EEG sensor are 50 Hz and 512 Hz, respectively. Then, the walking-motion data acquisition platform is constructed, in which all sensors are connected with the host computer through wireless bluetooth technology, as shown in Fig. 2.

VR-based media incentive is a new application of emotional stimulation. The media incentive is to maximize and effectively stimulate emotions through materials such as pictures, videos, music, and virtual environments. VR mainly provides virtual environment and augmented reality, that imitates the experience in various scenes and helps to stimulate real emotions. It can self-proactive regulation of brain activity to enhance emotional and cognitive processes. Kim et al.[25] used the VR environment and a pad-mounted pressure sensor to analyze the relationship between gait and emotional state. So we adopt the media incentive method based on VR. The VR-HMD equipment (iQIYI Inc., CHN) has built-in optical lenses with a field of view of 86°, which needs to be used with mobile smart phones and controlled by the handle. Subjects use VR-HMD equipment to watch VR videos during walking, and collect three kinds of walking-motion emotional data: neutral, happy, and fear.

Table 1 Main parameters of inertial sensor.

Component	Range	Resolution
Accelerometer	±16g	0.0005g/LSB
Gyroscope	±2000°/s	0.061°/s/LSB
Magnetometer	±0.0002 T	6.67×10^{-9} T/LSB

Note: LSB is the abbreviation of the least significant bit.

A total of 16 healthy subjects are recruited throughout the school to collect walking-motion emotional data, with an equal ratio of males and females. All subjects are mentally healthy, with no history of surgery or disease. To avoid any ambient noise, we select a quiet indoor site for data collection. After arriving at the experimental site, all subjects are told to wear EEG sensor and inertial sensors on the head and thighs respectively in advance, and then walk back and forth in the room according to their own habits. In this paper, inertial sensors and EEG sensors continuously track the subjects' movements and brain waves, thereby generating time series for further analysis. First, the subjects are familiar with the indoor environment, and then they are asked to walk for two minutes to collect walking-motion data in a neutral emotional condition. Second, subjects wear a powered-on VR-HMD device and select and play VR videos from an online VR video library according to their preferences to induce pleasure. The happy VR video content is that players get close to small animals and interact with them such as feeding and playing. After watching the VR video in advance to fully immerse

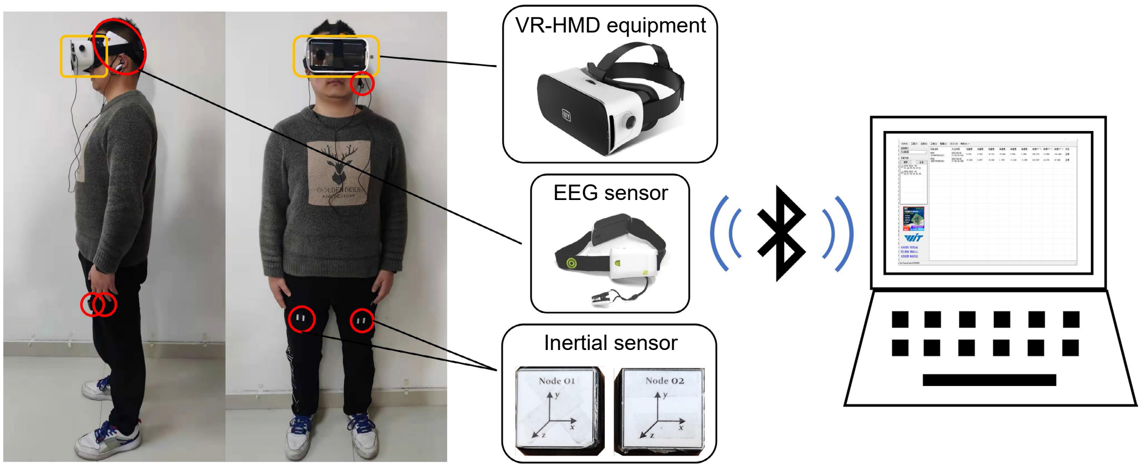


Fig. 2 Built walking-motion emotional data acquisition platform.

them, the subjects are asked to walk for two minutes, that is, to walk happily as expected. Third, to prevent the deviation caused by happy emotions to the greatest extent, the walking-motion data collection in fear emotions is carried out after an interval of half an hour. Likewise, subjects are asked to walk for two minutes after watching a frightening VR video. The fear VR video content is to travel through multiple haunted houses, with ghosts randomly teleporting behind the walls and corridors around the player. To avoid the cold start effect of emotions, we only select the emotional data of the subjects in the last minute and a half.

2.2 Data denoising

There is inevitable noise interference in the raw put data, involving electromagnetic interference from inside the device and packet loss in data transmission[26]. The vibrations generated when the body collides with the sensor can also interfere with the authenticity of the data. The original denoising method avoids the unnecessary additional activities in the experiment, but unintentional motion always causes noise. Therefore, further denoising is needed. Commonly used methods include moving average filtering, Kalman filterings, and wavelet denoising[27].

Kalman filtering is used for data preprocessing of inertial sensors, especially accelerometers, gyroscopes, and magnetometer[28]. It estimates the state of the target, establishes the state equation and combines the observation equation to process the inertial signal, finally realizes the accurate estimation of the information of the target at the next moment. This optimal estimation process is also a filtering process.

EEG signals are weak and easily affected by the power frequency interference of equipment and human physiological activities such as eye movement and breathing. In this paper, the threshold wavelet denoising method and moving average filter are used to preprocess the EEG signals. Four-layer decomposition Daubechies (Db) discrete wavelet is performed on the EEG signal. The effect of the hard threshold function is remarkable, but the reconstructed signal oscillates and is not smooth. So the moving average filter is used for smoothing, the equation is as follows:

$$y(n) = \frac{1}{l}\sum_{i=0}^{l-1} x(n-i) \tag{1}$$

where the input is $x(n)$, the output is $y(n)$, and l is the length of $x(n)$.

2.3 Data segmentation

Walking-motion is a cyclically repeated process, and it is necessary to divide the entire time series into sub-time series, which is to increase the number of samples to achieve more refined emotion recognition. This paper uses the sliding window method, as shown in Fig. 3. The selection of the width of the window affects the overall recognition effect of the system. If the window width is too large, multiple walking motions are contained within the window, otherwise, there is not enough data to represent the samples. We choose the number of data points collected within two seconds as the window width. And since the data between adjacent windows are related, we set 50% overlap between the two windows to avoid information loss. After segmentation, the dataset in this paper is a set M of T samples, $M = \{|X_1, y_1|, |X_2, y_2|, |X_t, y_t|, \ldots, |X_T, y_T|\}$, and the t-th sample is represented as

$$X_t = [S_{t,1},\ S_{t,2},\ S_{t,n}, \ldots,\ S_{t,N}] \tag{2}$$

where N is the total number of sensors worn by each subject, $S_{t,n}$ is the sample set of timing signals from the n-th sensor, and y_t is the manually labeled ground truth of X_t.

Since the inertial signal and the EEG signal have different channel numbers, the EEG signal is shaped to approximately match the length and channel number of the inertial signal. According to research[29], it is realized by converting a single-channel EEG signal into a 16-channel signal. Finally, we can obtain inertial samples with a size of 100×9 and EEG samples with a size of 64×16.

2.4 Data representation

For studying multiple data representations in emotion

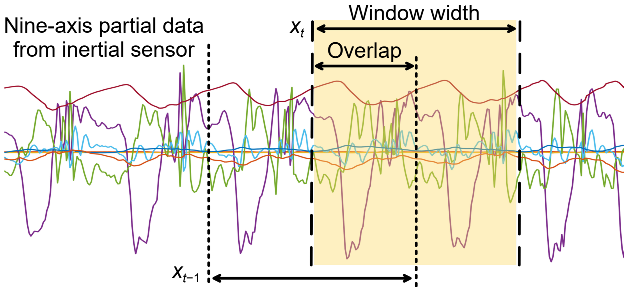


Fig. 3 Schematic diagram of data segmentation.

recognition and increasing the diversity of datasets, this paper converts EEG and inertial signals into virtual emotion images through DWT. EEG and inertial signals are typical non-stationary signals with strong randomness and their amplitude and frequency change with time. Fast Fourier transform (FFT) and discrete cosine transform (DCT) have a good solution to frequency domain transformation, but they are not suitable for the application of time domain transformation. Wavelet transform preserves the temporal and spatial information of frequencies and is the best candidate for the analysis and representation of these features in EEG and inertial signals.

DWT analyzes signals on different frequency bands with different resolutions through two mutual filters that decompose the signal into approximation and detail coefficients. The coefficients represent the low-frequency and high-frequency information of the signals, respectively. The wavelet coefficients are obtained by inner product of the input signal with the mother wavelet function and the scaling function, which can be computed by

$$
\begin{aligned}
w_\alpha[k,h] &= \frac{1}{\sqrt{2^k}}\sum_i x(i)\alpha\left(\frac{i-h2^k}{2^k}\right),\\
w_\beta[k,h] &= \frac{1}{\sqrt{2^k}}\sum_i x(i)\beta\left(\frac{i-h2^k}{2^k}\right)
\end{aligned}
\tag{3}
$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are mutually orthogonal basis functions, also called mother wavelets in discrete wavelet decomposition, k is the scale factor, and h is the displacement factor. After comparing different mother wavelets including Haar, Daubechies, Symlets, reverse biorthogonal, and biorthogonal, we consider 2.2-order reverse biorthogonal (rbio2.2) wavelet for our proposed method. Section 4.1 describes the details of the experiments.

Figure 4 shows the approximation coefficients (cA) and detail coefficients (cD) of the rbio2.2 wavelet obtained from the x-axis acceleration signal. N-level decomposition of discrete wavelet transform can get $N+1$ coefficients (N approximate coefficients and one detail coefficient), which can accurately represent the signal. Half of the samples can be eliminated according to the Nyquist theorem, which constitutes a level of decomposition, and this operation keeps increasing the frequency resolution. The length of the coefficients is approximately halved on a continuous level and can be calculated as

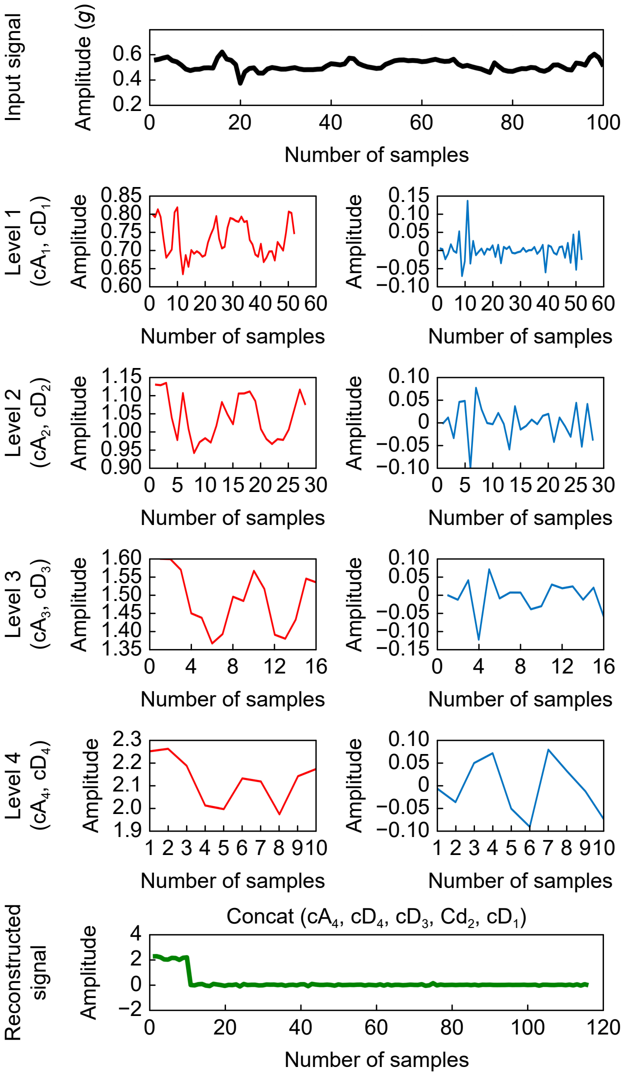


Fig. 4 Wavelet decomposition of DWT at the fourth decomposition level of the *x*-axis acceleration signal using rbio2.2 as mother wavelet.

$$
L = \text{floor}\left(\frac{l-1}{2}\right) + l_f \tag{4}
$$

where floor$(\cdot)$ is the round-down function, l is the length of the input signal, and l_f is half the length of the filter. The reconstructed signal is composed of detail coefficients decomposed at each layer and approximate coefficients of the last layer.

The input data need to be converted as formatted tensors in CNN, which is represented as $h \times w \times c$, where h, w, and c are height, width, and number of channels, respectively. The matrix obtained after representation from inertial and EEG sensor data is two-dimensional with a channel number of 1, and its form corresponds exactly to the grayscale image. The inertial signal subsequences are segmented from the inertial signals and normalized. Then, the sampling points in each subsequence after discrete wavelet transform are corresponding to the gray value to obtain images.

2.5 Attention-based CNN fusion model

The attention-based CNN structure of our proposed emotion recognition method includes a feature extraction module and an attention module as shown in Fig. 5. We use parallel and serial pure convolutional layers to form the backbone for extracting deep features, thus alleviating the structural complexity. After obtaining the multi-dimensional feature map, channel attention is used to focus on the salient parts and seek the most critical parts.

2.5.1 CNN architecture

CNN architecture mainly consists of convolutional layers and pooling layers. The convolution layer is the core of the CNN, and the convolution kernel moves in the input data matrix in the form of a sliding window, and the distance of each movement is called stride. The number of convolution kernels determines the number of channels in the output matrix. In the pooling layer, the dimension of the feature map can be reduced by adjusting the size and stride of the convolution kernel, thereby reducing the number of network parameters.

After DWT signal representation, we have prepared the input format for CNN. Here are T samples $\left\{\left|X_1^w, y_1\right|, \left|X_2^w, y_2\right|, \left|X_t^w, y_t\right|, \ldots, \left|X_T^w, y_T\right|\right\}$, and the t sample is represented as

$$X_t^w = \left[S_{t,1}^w,\ S_{t,2}^w,\ S_{t,n}^w, \ldots,\ S_{t,N}^w\right] \tag{5}$$

for each input $S_{t,n}^w$, CNN uses a two-dimensional convolution operation to extract features layer by layer. The convolution values mapped at the position $(i,\ j)$ of the d-layer feature map as

$$F_{i,j}^d = \left(F^{d-1} \times K\right)_{i,j} = \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} F_{i+p,\ j+p}^{d-1} K_{p,q} \tag{6}$$

where d is the convolution layer index, $K_{p,q}$ is the value of the kernel function at position $(p,\ q)$, P and Q are the height and width of the convolution kernel, respectively.

To extract and optimize the correlation features between the multi-channel signals of each sensor, we design a CNN structure for EEG and inertial signals as shown in Fig. 6. The principle is to use multiple convolution kernels of different sizes to extract the features of different signal channels in parallel, and then two convolution layers are used to optimize the extracted features. By designing network parameters, the concatenated feature map is arranged according to the channel dimension, which needs to ensure that the length and width of the feature maps are consistent as shown in Fig. 7. The proposed CNN structure adds batch normalization (BN) layers to speed up network training and convergence. Each convolutional layer performs two-dimensional convolution, followed by a BN layer and a ReLU activation function. First, there are 64 convolution kernels of size 1×3, 5×5, and 7×7 to detect the features of different signal channels. Then continue to use two convolutional layers to optimize the features, each convolutional layer is followed by a pooling layer, and the size of the built-in pooling kernel is 2×2. After the second max-pooling layer, the feature matrix is flattened into a feature vector. The specific parameters of CNN structure are shown in Table 2.

2.5.2 Attention mechanism

The main function of CNN is to extract features, and the task of feature optimization is handed over to the attention mechanism. We use an attention mechanism to learn the correlations in channel dimensions of the feature matrix. According to the learned correlation, the channel weight matrix is obtained and multiplied with the original feature matrix to enhance the part with key information. The specific structure of the proposed attention mechanism is shown in Fig. 8.

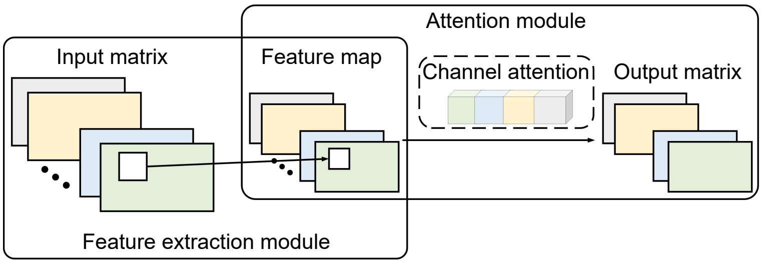


Fig. 5 Attention-based CNN structure containing feature extraction module and attention module.

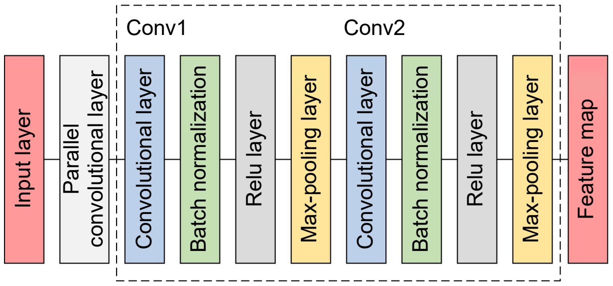


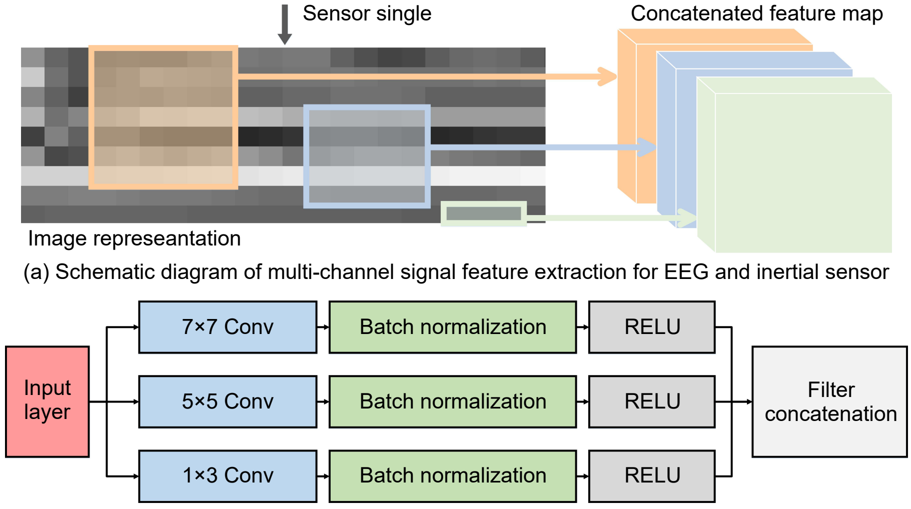
Fig. 6 CNN architecture for EEG and inertial signals.

The feature map F after CNN can be expressed as $F=\{F_1,F_2,\ldots,F_c\}$, where c is the number of channels of F. Firstly, the average-pooling and max-pooling of F_c are used to extract the global spatial information. Average-pooling and max-pooling calculate the average and maximum values of elements in the pooling window, respectively. Both are used for feature dimension reduction, they are defined as

$$\text{out}_{\text{max-pooling}}=\max\left[x_1,x_2,\ldots,x_n\right] \tag{7}$$

$$\text{out}_{\text{avg-pooling}}=\text{mean}\left[x_1,x_2,\ldots,x_n\right] \tag{8}$$

where x is the input data, so the maximum pooling feature $F_{\max}$ and the average pooling feature F_{avg} can be obtained. Then, two-layer one-dimensional convolution $\text{Conv1d}(\cdot)$ is used to realize the information interaction between channels. The number of convolution kernels in the two layers is $c\times r$ and c,



(a) Schematic diagram of multi-channel signal feature extraction for EEG and inertial sensor

(b) Parallel structure of multiple convolutional layers

Fig. 7 Illustration of the feature extraction module.

Table 2 Parameters of CNN structure.

Layer name	Number of kernels	Kernel size	Stride	Padding
Parallel Conv	[64, 64, 64]	[(1, 3), 5, 7]	1	[(0, 1), 2, 3]
Conv1	128	3	1	0
Max-pooling 1	None	2	2	0
Conv2	64	3	1	0
Max-pooling 2	None	2	2	0

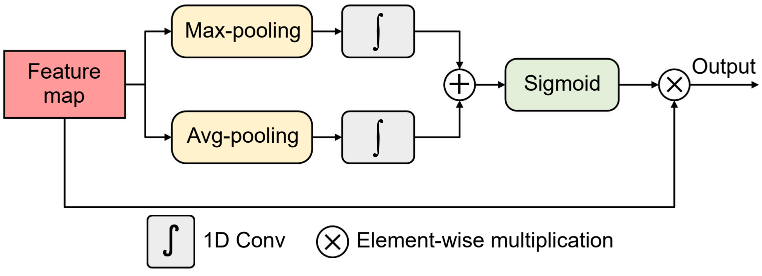


Fig. 8 Structure of the channel attention mechanism.

respectively, where r is the scale factor in the range of 0 to 1. After that, two vectors are added according to the corresponding positions of the elements to obtain c-dimensional vector F_c is obtained, as follows:

$$F_c = \text{Conv1d}\left(F_{\text{avg}}\right) + \text{Conv1d}(F_{\max}) = [F_{\text{avg},1}, F_{\text{avg},2}, \ldots, F_{\text{avg},c}] + [F_{\max,1}, F_{\max,2}, \ldots, F_{\max,c}] = [F_{c,1}, F_{c,2}, \ldots, F_{c,c}] \tag{9}$$

Finally, the sigmiod activation function is used to obtain the c-dimensional vector, and the value of each element is the weight corresponding to each channel belonging to [0, 1]. The output of sigmiod function is

$$\text{out}_{\text{score}} = \frac{\exp(F_c)}{\exp(F_c)+1} \tag{10}$$

and the mapping structure of channel attention is as follows:

$$F_{\text{atten}} = F \odot \text{out}_{\text{score}} \tag{11}$$

where $\odot$ represents the multiplication of elements one by one.

2.5.3 Fusion

Multi-modal data fusion in emotion recognition is performed at different levels and divided into input-level fusion, feature-level fusion, and decision-level fusion. Input-level fusion is the fusion of the input data phase, preserving as much information as possible. Feature-level fusion is to perform fusion after feature extraction, which greatly retains the original information while realizing data compression. Decision-level fusion is integrated at the classification and discrimination level, with a high fault tolerance rate, and higher-level decision-making can be made according to application requirements.

Feature-level fusion. Input-level data fusion is considered for processing the processed two inertial sensor signals. The inertial signals of each emotion are fused using concatenation, resulting in an image of size 100×18. Feature-level fusion can comprehensively utilize multimodal features to achieve complementary advantages, which improves the robustness and accuracy of the system. After extracting features from the EEG and inertial sensors, we independently compute the feature vectors obtained from each modality after passing through the flattening layer. And the feature vectors corresponding to each emotion are concatenated to obtain a new high-dimensional feature vector. We can connect the feature matrix by

$$F_{\text{fusion}} = \text{concat}(F_{\text{inertial}}, F_{\text{EEG}}) \tag{12}$$

where F_{fusion} is the fused feature, and concat$(\cdot)$ represents the connection operation. Moreover, it is necessary to balance the new features, which means that the various types of features stitched together have the same numerical scale. We apply Min-Max normalization to the resultant features and map to the range of [0, 1] according to the following:

$$y = \frac{x_{\text{in}} - \min(x_{\text{in}})}{\max(x_{\text{in}}) - \min(x_{\text{in}})} \tag{13}$$

where y is the normalized data and x_{in} is the original data.

Decision-level fusion. The prediction results obtained by individual recognition are usually given equal weights, and then the emotional prediction is obtained through specific fusion strategies. However, this lacks consideration of the correlation between the multi-modal data. In addition, strong classification models are easily overwhelmed by some weak classification models. We propose a decision fusion algorithm considering a single modality model and a feature fusion model. The algorithm uses Critic method to assign weights pairs to the predicted labels and fuse them according to the majority voting strategy.

The Critic method is to comprehensively measure the objective weight of the indicators based on the contrast strength of the evaluation indicators and the conflict between the indicators. The fusion strategy refers to the research[30]. Assuming that there is a class a emotion, the sample can be expressed as $X = \{x_1, x_2, \ldots, x_p, \ldots, x_n\}$. We assume that there are b classification models. For the test sample x_p, the classification result using b models is expressed as $M = \{m_{p1}, m_{p2}, \ldots, m_{pq}, \ldots, m_{pb}\}$, where m_{pq} represents the classification result of the p-th test sample under the q-th classification model. Then, the Critic weight majority voting algorithm can be represented by

$$\Gamma_e = \mathop{\arg\max}_{e \in (1,2,\ldots,a)} \sum_{i=1}^{b} \gamma_{ei} \theta_i(e) \tag{14}$$

where

$$\theta_i(e) = \begin{cases} 1, & x_p \in \text{emotion } e; \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

r_{gh} $(1 \leqslant g \leqslant a, 1 \leqslant h \leqslant b)$ represents the recall of the a-th class emotion under the b-th classification model, the valuation matrix R as

$$R=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1b} \\ r_{21} & r_{22} & \cdots & r_{2b} \\ \vdots & \vdots & \cdots & \vdots \\ r_{a1} & r_{a2} & \cdots & r_{ab} \end{pmatrix} \tag{16}$$

Then, the index variability S_h and index conflict C_h of the b-th classification model are expressed in the form of standard deviation and correlation coefficient, respectively, as follows:

$$S_h=\sqrt{\frac{\sum\limits_{i=1}^{a}(r_{ih}-\bar{r}_h)^2}{a-1}} \tag{17}$$

$$C_h=\sum_{j=1}^{b}1-c_{jh} \tag{18}$$

$$\bar{r}_h=\frac{1}{a}\sum_{i=1}^{a}r_{ih} \tag{19}$$

where c_{jh} represents the correlation coefficient between the j-th classification model and the h-th classification model.

The greater the amount of information Y, the greater the role of the j-th evaluation index in the entire evaluation index system, and more weights should be assigned to it.

$$Y_h=S_h\times C_h \tag{20}$$

Finally, the weight γ_h of the h-th classification model for the e-th emotion can be obtained as

$$\gamma_h=\frac{Y_h}{\sum\limits_{i=1}^{b}Y_i} \tag{21}$$

3 Classification Module and Loss Function

After obtaining emotional features, a classification module is designed to map high-dimensional features into a low-dimensional space for emotion recognition. First, the features are flattened to obtain a one-dimensional vector, which is then mapped through a multi-layer neural network. The output of the j-th neuron as

$$y_j=f\left(\theta+\sum_{i=1}^{n}\omega_{ij}x_i\right) \tag{22}$$

where x_i is the i-th input, ω_{ij} is the weight of the i-th input of the j-th neuron, and θ is the bias.

To significantly reduce the phenomenon of overfitting in machine learning, dropout is used as a trick for training deep neural networks. To significantly reduce the phenomenon of overfitting in machine learning, dropout is used as a trick for training deep neural networks by ignoring a certain proportion of feature detectors (hidden layer node value is 0) in each training batch. The last fully connected layer maps the dimension to the number of emotion types. The detection and recognition of each emotion are achieved based on the output scores generated by the softmax layer. The softmax score of the i-th emotion as

$$\text{softmax}(y_i)=\frac{e^{y_i}}{\sum\limits_{j=1}^{n}e^{y_j}} \tag{23}$$

During the model training process, the learning rate and batch size are set to 0.001 and 8, respectively. Gayscale images are selected as the input to the neural network, and the average error is calculated by comparing the actual prediction results with the actual output results. For multi-classification tasks, we use the cross-entropy loss function to measure the deviation between real values and predicted values:

$$\text{Loss}=-\sum_{i=1}^{n}p(x_i)\log q(x_i) \tag{24}$$

where $p(x_i)$ is the true value of the input x_i and $q(x_i)$ is the predicted value.

4 Experiment

All emotional samples are divided into ten parts, eight of which are used as training data to train the proposed model using leave-one-out cross-validation, and the remaining two parts are used as the test data of the trained model. Then, the training data are divided into five parts, four of which are used for training and the rest for validation. This process is repeated five times until every data is used for validation. The recognition performance of the three different emotions is evaluated according to four evaluation indicators: accuracy (ACC), precision (PRE), recall (REC), and F_1 score (F_1).

4.1 Evaluation of details about mother wavelets and decomposition levels in DWT

A comparative analysis was carried out on the most effective mother wavelet type in the discrete decomposition of EEG and inertial signals collected during walking in three emotions. Wavelets with high support tend to have difficult to detect the closely spaced features used to recognize different emotional states during walking, and the support of wavelets should be small enough to separate the features of interest. Therefore, the mother wavelets selected for comparison include Haar, 2-order Daubechies (Db2), 2-order Symlets (Sym2), 2.2-order reverse biorthogonal (rbio2.2), and 2.2-order biorthogonal (bior2.2) for the recognition of three different emotions, and the comparison results are shown in Table 3. For each emotional data after wavelet decomposition and reconstruction, the attention-based CNN structure mentioned in Section 2.5 is followed. Firstly, the specific effect of the EEG signal after DWT as the input of the attention-based CNN structure on the recognition of three emotions during walking is analyzed. It can be seen from Table 3 that the rbio2.2 mother wavelet achieves the best performance, with accuracy, precision, recall, and F_1 score of 84.39%, 85.48%, 84.36%, and 84.74%, respectively. The performance of the models using Haar mother wavelet decomposition is relatively lower than other wavelets, with accuracy, precision, recall, and F_1 score of 77.22%, 78.82%, 77.25%, and 77.78%, respectively. The main reason is that Haar wavelet transform is not sensitive enough to the local feature of EEG signals, which may lead to the loss of emotional information. Then the influence of inertial signal on the recognition effect of different emotional states during walking is analyzed, rbio2.2 and Haar have almost the same performance. The accuracy, precision, recall, and F_1 score of Haar are 97.05%, 97.15%, 97.20%, and 97.18%, respectively, and the accuracy, precision, recall, and F_1 score of rbio2.2 are 97.05%, 97.10%, 97.20%, and 97.15%, respectively. The results show that rbio2.2 and Haar wavelet can accurately extract the local emotional features of low frequency and high frequency in the inertial signal, and retain the useful emotional information in the original signal. The comparison of model accuracy performance for wavelet decomposition with different types of mother wavelets is shown in Fig. 9. It can be seen that the model using rbio2.2 wavelet has the highest accuracy compared to other mother wavelets for EEG and inertial signals. So this paper uses rbio2.2 as the mother wavelet for discrete decomposition of EEG and inertial signals.

Table 3 Performance of EEG and inertial signals on wavelet decomposition with different types of mother wavelets.

Input signal	Wavelet name	ACC (%)	PRE (%)	REC (%)	F_1 (%)
EEG signal	Haar	77.22	78.82	77.25	77.78
	Db2	82.28	82.58	82.43	82.50
	Sym2	78.48	79.50	79.00	79.06
	bior2.2	81.43	82.19	81.52	81.79
	rbio2.2	**84.39**	**85.48**	**84.36**	**84.74**
Inertial signal	Haar	**97.05**	**97.15**	**97.20**	**97.18**
	Db2	94.94	95.33	95.00	95.14
	Sym2	96.20	96.49	96.24	96.35
	bior2.2	96.20	96.29	96.40	96.34
	rbio2.2	**97.05**	97.10	**97.20**	97.15

This paper also compares the evaluation of emotion recognition performance of rbio2.2 mother wavelet decomposition level. The increase and decrease of the decomposition layer has an impact on the recognition of emotions during walking. The increase of the decomposition layer leads to signal distortion, and the decrease of the decomposition layer cause the emotional features in the signal to not be fully characterized. Level 0 refers to the performance of an attention-based CNN model trained using signals without DWT to recognize emotions during walking. Figure 10 shows the recognition accuracy of three emotions with different wavelet decomposition levels. The DWT of rbio2.2 mother wavelet with a 2-level decomposition for EEG and inertial signal has the best recognition effect of three emotions. The analysis in Fig. 10 shows that with the increase of the decomposition degree, the recognition accuracy of different emotions tends to increase, and the accuracy performance reach 97.05% and 84.39% in the 2-level decomposition for inertia and EEG signals,

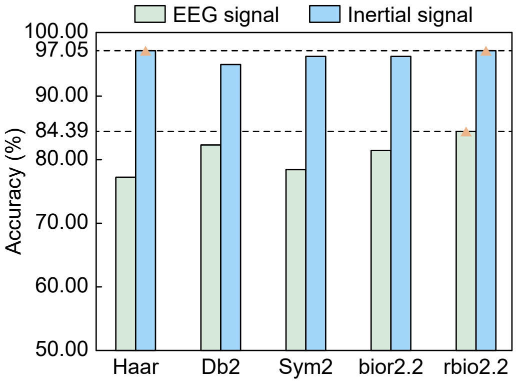


Fig. 9 Comparison of the accuracy performance for EEG and inertial signals in wavelet decomposition of different types of mother wavelets.

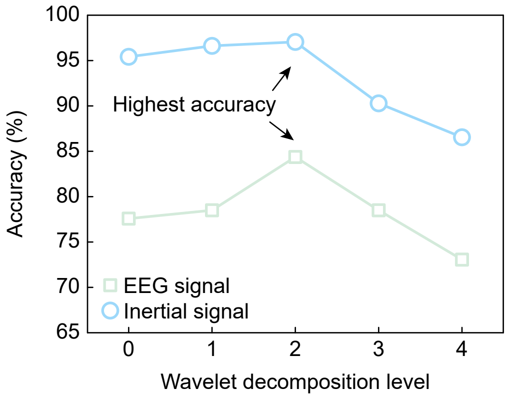


Fig. 10 Comparison of the accuracy performance for EEG and inertial signals in wavelet decomposition of different levels.

respectively. Then the performance of the accuracy tends to decrease with the increasing degree of decomposition. The reason behind the downgrade can be traced back to the fact that an increase in level leads to loss of emotional information. In addition, the increase of the decomposition level increases the complexity of the algorithm. The specific performance of the recognition of three emotions with different levels of wavelet decomposition is shown in Table 4.

4.2 Evaluation of different data representation methods

To further evaluate the effect of the data representation on the recognition of different emotions during walking, we conducted an experimental analysis. The proposed DWT-based data representation is compared with FFT and DCT, which have gained attention for their suitability for signal processing in deep learning. Table 5 shows the emotion recognition performance of different data representation methods. For EEG and inertial signals, the proposed method for signal representation using DWT achieves the highest recognition accuracy for three emotions. In the EEG signal representation, the overall performance of DCT is better than FFT, and its accuracy, precision, recall, and F_1 score are 75.53%, 77.49%, 75.36%, and 76.06%, respectively, while the accuracy, precision, recall, and F_1 score of FFT are 69.20%, 70.75%, 69.56%, and 69.78%, respectively. The performance of FFT is better than that of DCT in the representation of inertial signals, with accuracy, precision, recall, and F_1 score of 90.30%, 90.56%, 90.65%, and 90.57%, respectively, and the accuracy, precision, recall, and F_1 score of DCT are 85.65%, 86.68%, 85.86%, and 86.19%, respectively. Figure 11 intuitively shows the specific performance of the evaluation indicators of the three data representation methods in the overall emotion recognition system. The reason why we choose DWT as the representation of EEG and inertial signals is that DCT and FFT can only convert the emotional signal from the time domain to the frequency domain, which shows the corresponding amplitude at different frequencies. DWT can not only examine the frequency domain feature of the emotional signal in the local time domain process, but also examine the time domain characteristics of the local frequency domain process. Therefore, DWT enables well transform and process non-stationary signals such as EEG and inertial signals in emotion recognition during walking.

Table 4 Performance of EEG and inertial signals on wavelet decomposition with different levels.

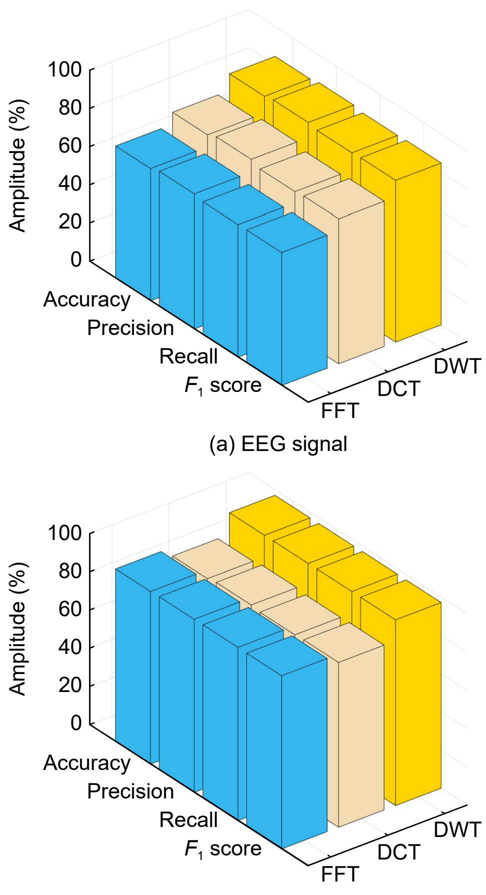
Input signal	Decomposition level	ACC (%)	PRE (%)	REC (%)	F_1 (%)
EEG signal (rbio2.2)	0	77.64	78.04	77.82	77.55
	1	78.48	79.68	78.74	78.91
	2	**84.39**	**85.48**	**84.36**	**84.74**
	3	78.48	78.78	78.90	78.84
	4	73.00	74.32	73.23	73.67
Inertial signal (rbio2.2)	0	95.36	95.56	95.71	95.57
	1	96.62	96.79	96.95	96.78
	2	**97.05**	**97.10**	**97.20**	**97.15**
	3	90.30	90.66	90.63	90.64
	4	86.50	86.86	86.82	86.73

Table 5 Performance comparison of different data representation methods.

Input signal	Method	ACC (%)	PRE (%)	REC (%)	F_1 (%)
EEG signal	FFT	69.20	70.75	69.56	69.78
	DCT	75.53	77.49	75.36	76.06
	DWT	**84.39**	**85.48**	**84.36**	**84.74**
Inertial signal	FFT	90.30	90.56	90.65	90.57
	DCT	85.65	86.68	85.86	86.19
	DWT	**97.05**	**97.10**	**97.20**	**97.15**

4.3 Evaluation of the effectiveness of the attention mechanism

Just as humans can allocate different attention to different places when doing complex work, the attention mechanism can give neural networks the ability to focus on their feature maps. To verify the effectiveness of the CNN structure based on the attention mechanism, we design a channel attention mechanism to assign weights and divide the importance of the channels of the obtained feature maps. For the CNN structure without attention mechanism, we directly input the obtained feature map to the classification module after the flattening layer. Table 6 shows the specific performance evaluation of the attention mechanism. For EEG signal, the accuracy, precision, recall, and F_1 score of the emotion recognition system are improved by 4.64%, 5.25%, 4.08%, and 4.65%, respectively. For inertial signal, the accuracy, precision, recall, and F_1 score of the emotion recognition system are improved by 3.80%, 3.56%, 3.72%, and 3.66%, respectively. For feature fusion, the accuracy, precision, recall, and F_1 score of the emotion recognition system are improved by 1.69%, 1.66%, 1.65%, and 1.66%, respectively. Compared with the CNN structure without the attention mechanism, the accuracy of the attention-based CNN structure is improved by a maximum of 5.25% and a minimum of 1.65%. Figure 12 shows the confusion matrix of the recognition results of the samples in different emotions based on the attention-based CNN architecture. Each column and each row of the confusion matrix represent the predicted emotional category and the real emotional category of the sample, respectively, where each data is the proportion of the number of specific emotional samples to the total number of samples in the test set.



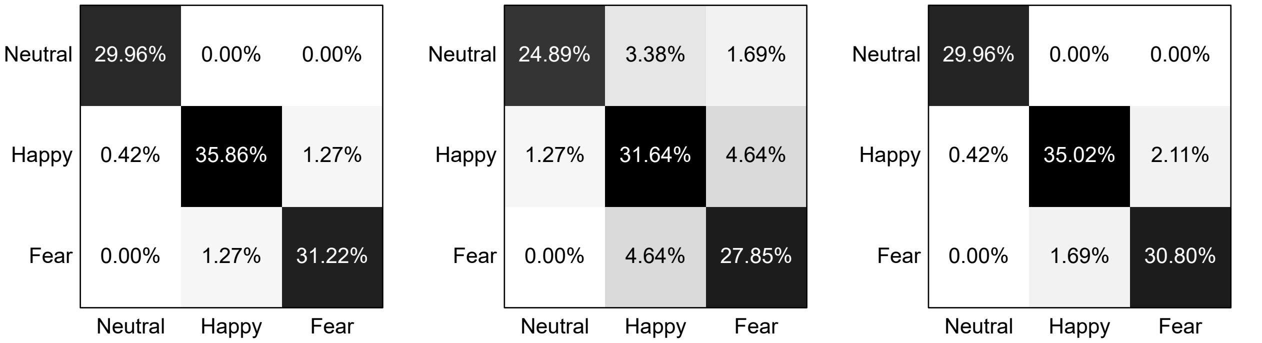
(a) EEG signal

(b) Inertia signal

Fig. 11 Specific performance comparison of the three data representation methods.

Table 6 Performance evaluation of the effectiveness of the attention mechanism.

Method	Input	ACC (%)	PRE (%)	REC (%)	F_1 (%)
Without attention mechanism	EEG signal	79.75	80.23	80.28	80.09
	Inertial signal	93.25	93.54	93.48	93.49
	Feature fusion	94.09	94.21	94.37	94.28
With attention mechanism	EEG signal	84.39	85.48	84.36	84.74
	Inertial signal	97.05	97.10	97.20	97.15
	Feature fusion	95.78	95.87	96.02	95.94



(a) Inertia signal (accuracy: 97.05%)

(b) EEG signal (accuracy: 84.39%)

(c) Feature fusion (accuracy: 95.78%)

Fig. 12 Confusion matrices for the attention-based CNN architecture.

4.4 Evaluating the decision fusion algorithm of EEG and inertial signals

We consider two aspects in decision-level fusion. On one hand, feature fusion considers the correlation between EEG and inertial signals. On the other hand, the decision-making ability of each model participating in the decision fusion is evaluated, and the Critic method is used to add weights to the majority voting mechanism. To evaluate the performance of the proposed decision fusion algorithm on the performance of emotion recognition system, we compare with majority voting and decision fusion algorithm without feature fusion, and the results are shown in Table 7. The decision fusion algorithm does not have feature fusion, which means that only EEG and inertial signals are fused. Its accuracy, precision, recall, and F_1 score are 97.47%, 97.54%, 97.58%, and 97.55%, respectively. The decision fusion algorithm proposed in this paper considers feature fusion, and the accuracy, precision, recall, and F_1 score are 98.73%, 98.77%, 98.82%, and 98.79%, respectively. Therefore, feature fusion plays a role in supplementing emotional information and reducing emotional prediction errors. For majority voting, it performs moderately, with accuracy, precision, recall, and F_1 score of 97.89%, 98.01%, 97.95%, and 97.98%, respectively. Figure 13 shows the confusion matrix comparing the three decision fusion methods. Compared with the other two methods, the proposed decision fusion algorithm improves the recognition of happiness and fear emotions, and the number of samples misclassified as other emotions is significantly reduced. The proposed decision fusion algorithm solves the problem of equal opportunity of emotion recognition system in most voting, and improves the accuracy of emotion recognition during walking to a certain extent. In multi-modal fusion, the decision fusion algorithm considers the correlation between the two input modes of inertia and EEG signals, which is reflected in feature fusion. In terms of accuracy, the proposed fusion algorithm is significantly better than the majority voting and decision fusion algorithms without feature fusion. In general, the decision fusion algorithm proposed in this paper is effective for improving the recognition performance of three emotions during walking.

Table 7 Performance evaluation of the effectiveness of the proposed decision fusion algorithm.

(%)

Method	ACC	PRE	REC	F_1
Proposed decision fusion	**98.73**	**98.77**	**98.82**	**98.79**
Without feature fusion	97.47	97.54	97.58	97.55
Majority voting	97.89	98.01	97.95	97.98

4.5 Two-category classification of emotions during walking

The purpose of the two-category classification is to study the prediction accuracy of the model proposed in this paper for each of the two emotions (3 combinations in total), which are a set of 3 basic emotions, including neutral and happy, neutral and fear, and happy and fear. The comparison results of three kinds of emotion combination recognition are shown in Table 8. Neutral-happy has the highest classification performance with accuracy, precision, recall, and F_1 score of 99.38%, 99.31%, 99.44%, and 99.37%, respectively. Happy-fear is slightly inferior to the other two, its accuracy, precision, recall, and F_1 score are 93.37%, 93.34%, 93.56%, and 93.36%, respectively. It can be seen that the walking motion under neutral emotion is distinct and easy to identify from those under happiness and fear.

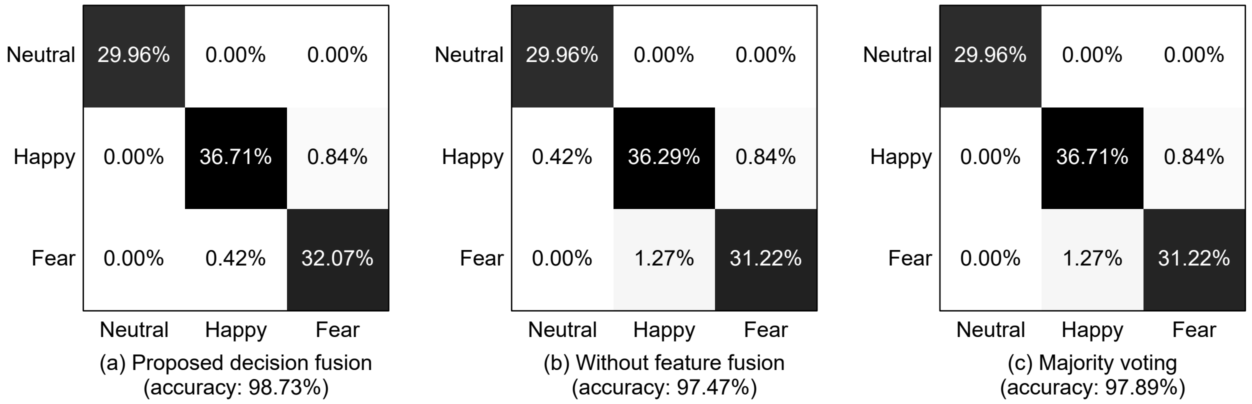


Fig. 13 Confusion matrix comparison of three decision fusion methods.

Table 8 Comparison of the classification performance of three types of emotion combinations.

(%)

Emotion	ACC	PRE	REC	F_1
Neutral-happy	**99.38**	**99.31**	**99.44**	**99.37**
Neutral-fear	98.65	98.73	98.59	98.64
Happy-fear	93.37	93.34	93.56	93.36

4.6 Comparison with other existing state-of-the-art works

This paper compares the performance of our proposed multi-modal emotion recognition method with three state-of-the-art works. The premise of the comparison of these methods is to use the data set proposed in this paper, and the division of training set, validation set, and test set remains unchanged. Table 9 summarizes the comparison. Zhang et al.[31] developed a physiological signal-based, mean-threshold, and decision-level fusion algorithm based on EEG and PERipheral physiological signals to explore emotion recognition. The algorithm first selects emotion-related features from the signals, and then uses classical classifiers such as Gaussian Naive Bayes, Linear Regression, and Support Vector Machine to establish a single-modal classification model. Finally, the decision-level fusion algorithm is used to integrate these into a new integrated classification model to improve the classification accuracy. Its accuracy, precision, recall, and F_1 score are 92.05%, 92.40%, 92.36%, and 92.38%, respectively. The main reason is that although the features extracted manually can intuitively reflect the correlation with emotions, they may ignore some useful functions and cannot be applied to different scenarios. Liu et al.[32] constructed an importance attention network with complementary modalities. Considering that there is a certain complementary relationship between the importance differences between multiple modes, they fuse the reconstructed features to obtain multi-modal features with good interactivity. And it achieves performance of 95.33%, 95.48%, 95.67%, and 95.40% for accuracy, precision, recall, and F_1 score, respectively. Xu et al.[33] proposed a bi-modal emotion recognition framework composed of parallel convolution (Pconv) module and attention-based bi-directional long short-term memory (BLSTM) module. Pconv module provides more effective representation capabilities to extract multi-dimensional social features using parallel methods, and attention-based BLSTM module strengthens the extraction of key information and maintains the correlation between information. Its accuracy, precision, recall, and F_1 score are 96.75%, 96.03%, 96.93%, and 96.48%, respectively. These two tasks complement each other by feature fusion of different data patterns, thereby ensuring robustness and improving the accuracy of emotion recognition, but lack flexibility and anti-interference compared to decision-level fusion. Overall, our proposed method achieves higher system performance than other methods, which is attributed to the attention-based CNN structure extracting the most discriminative features and decision fusion algorithm flexibly assigns label weights.

Table 9 Comparison with other existing state-of-the-art works on the dataset proposed in this paper.

(%)

Method	ACC	PRE	REC	F_1
Zhang et al.[31]	92.05	92.40	92.36	92.38
Liu et al.[32]	95.33	95.48	95.67	95.40
Xu et al.[33]	96.75	96.03	96.93	96.48
This paper	**98.73**	**98.77**	**98.82**	**98.79**

5 Conclusion

This paper proposes a method for recognizing emotions during walking using EEG and inertial signals. The subjects are accompanied by immersive emotions during walking through VR-based media incentives and complete the acquisition of multi-modal emotional data. We propose and evaluate the effectiveness of DWT-based data representation methods, and use them as input to the attention-based CNN structure to extract significant parts of relevant features. The proposed decision fusion algorithm combines the Critic method and majority voting strategy to fully consider the influence of EEG and inertial signals on emotion recognition. Compared with other state-of-the-art methods reimplemented on the dataset constructed in this paper, the proposed method has the highest accuracy of 98.73%. An interesting approach for future research involves the efficient combination of kinect depth cameras and wearable sensors. We also hope to explore a graph convolutional neural network model that can process human skeleton images to analyze the influence of walking joint trajectories on emotion recognition performance. At the same time, we are looking for a new method based on gait analysis, which can combine temporal features and spatial features according to the gait cycle to accurately, and stably recognize the emotion during walking.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61903170, 62173175, and 61877033), the Natural Science Foundation of Shandong Province (Nos. ZR2019BF045 and ZR2019MF021), and the Key Research and Development Project of Shandong Province of China (No.2019GGX101003).

References

[1] F. Y. N. Leung, J. Sin, C. Dawson, J. H. Ong, C. Zhao, A. Veić, and F. Liu, Emotion recognition across visual and auditory modalities in autism spectrum disorder: A systematic review and meta-analysis, *Dev. Rev.*, vol. 63, p. 101000, 2022.

[2] W. K. Ngai, H. Xie, D. Zou, and K. L. Chou, Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources, *Inf. Fusion*, vol. 77, pp. 107–117, 2022.

[3] P. Parada-Fernández, D. Herrero-Fernández, R. Jorge, and P. Comesaña, Wearing mask hinders emotion recognition, but enhances perception of attractiveness, *Pers. Individ. Differ.*, vol. 184, p. 111195, 2022.

[4] Y. Bhatia, A. H. Bari, G. J. Hsu, and M. Gavrilova, Motion capture sensor-based emotion recognition using a bi-modular sequential neural network, *Sensors*, vol. 22, no. 1, p. 403, 2022.

[5] S. Qiu, Z. Wang, H. Zhao, K. Qin, Z. Li, and H. Hu, Inertial/magnetic sensors based pedestrian dead reckoning by means of multi-sensor fusion, *Inf. Fusion*, vol. 39, pp. 108–119, 2018.

[6] H. Zhao, Z. Wang, S. Qiu, Y. Shen, L. Zhang, K. Tang, and G. Fortino, Heading drift reduction for foot-mounted inertial navigation system via multi-sensor fusion and dual-gait analysis, *IEEE Sens. J.*, vol. 19, no. 19, pp. 8514–8521, 2019.

[7] T. T. Pham and Y. S. Suh, Conditional generative adversarial network-based regression approach for walking distance estimation using waist-mounted inertial sensors, *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[8] D. Seckiner, X. Mallett, P. Maynard, D. Meuwly, and C. Roux, Forensic gait analysis—Morphometric assessment from surveillance footage, *Forensic Sci. Int.*, vol. 296, pp. 57–66, 2019.

[9] S. Pal, S. Mukhopadhyay, and N. Suryadevara, Development and progress in sensors and technologies for human emotion recognition, *Sensors*, vol. 21, no. 16, p. 5554, 2021.

[10] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges, *Inf. Fusion*, vol. 80, pp. 241–265, 2022.

[11] I. Mohino-Herranz, R. Gil-Pita, J. García-Gómez, M. Rosa-Zurera, and F. Seoane, A wrapper feature selection algorithm: An emotional assessment using physiological recordings from wearable sensors, *Sensors*, vol. 20, no. 1, p. 309, 2020.

[12] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al., Universals and cultural differences in the judgments of facial expressions of emotion, *J. Pers. Soc. Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.

[13] J. A. Russell, A circumplex model of affect, *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[14] J. A. Russell and A. Mehrabian, Distinguishing anger and anxiety in terms of emotional response factors, *J. Consult. Clin. Psychol.*, vol. 42, no. 1, pp. 79–83, 1974.

[15] M. A. Hashmi, Q. Riaz, M. Zeeshan, M. Shahzad, and M. M. Fraz, Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone, *IEEE Sens. J.*, vol. 20, no. 22, pp. 13511–13522, 2020.

[16] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu, Emotion recognition based on customized smart bracelet with built-in accelerometer, *PeerJ*, vol. 4, p. e2258, 2016.

[17] M. Aslan, CNN based efficient approach for emotion recognition, *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7335–7346, 2022.

[18] A. V. Atanassov, D. I. Pilev, F. N. Tomova, and V. D. Kuzmanova, Hybrid system for emotion recognition based on facial expressions and body gesture recognition, in *Proc. 2021 Int. Conf. Automatics and Informatics* (*ICAI*), Varna, Bulgaria, 2021, pp. 135–140.

[19] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM, *Appl. Soft Comput.*, vol. 100, p. 106954, 2021.

[20] T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, and N. Jiang, A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition, *Comput. Biol. Med.*, vol. 159, p. 106938, 2023.

[21] M. Borghetti, M. Serpelloni, E. Sardini, and O. Casas, Multisensor system for analyzing the thigh movement during walking, *IEEE Sens. J.*, vol. 17, no. 15, pp. 4953–4961, 2017.

[22] F. Y. Liang, F. Gao, and W. H. Liao, Synergy-based knee angle estimation using kinematics of thigh, *Gait Posture*, vol. 89, pp. 25–30, 2021.

[23] S. Wang, Y. Wang, D. Liu, Z. Zhang, W. Li, C. Liu, T. Du, X. Xiao, L. Song, H. Pang, et al., A robust and self-powered tilt sensor based on annular liquid-solid interfacing triboelectric nanogenerator for ship attitude sensing, *Sens. Actuat. A Phys.*, vol. 317, p. 112459, 2021.

[24] W. Zhang, Y. Liu, S. Zhang, T. Long, and J. Liang, Error fusion of hybrid neural networks for mechanical condition dynamic prediction, *Sensors*, vol. 21, no. 12, p. 4043, 2021.

[25] Y. Kim, J. Moon, N. J. Sung, and M. Hong, Correlation between selected gait variables and emotion using virtual

reality, *J. Ambient Intell. Humaniz. Comput.*, https://doi.org/10.1007/s12652-019-1456-2, 2019.

[26] I. H. López-Nava and A. Muñoz-Meléndez, Wearable inertial sensors for human motion analysis: A review, *IEEE Sens. J.*, vol. 16, no. 22, pp. 7821–7834, 2016.

[27] Z. Wang, M. Guo, and C. Zhao, Badminton stroke recognition based on body sensor networks, *IEEE Trans. Hum. Mach. Syst.*, vol. 46, no. 5, pp. 769–775, 2016.

[28] Y. Zhao, M. Guo, X. Sun, X. Chen, and F. Zhao, Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals, *IET Signal Process.*, vol. 17, no. 4, p. e12201, 2023.

[29] S. S. Bangaru, C. Wang, S. A. Busam, and F. Aghazadeh, ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors, *Autom. Constr.*, vol. 126, p. 103653, 2021.

[30] M. Guo, Z. Wang, N. Yang, Z. Li, and T. An, A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors, *IEEE Trans. Hum. Mach. Syst.*, vol. 49, no. 1, pp. 105–111, 2019.

[31] Q. Zhang, H. Zhang, K. Zhou, and L. Zhang, Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition, *Tsinghua Science and Technology*, vol. 28, no. 4, pp. 673–685, 2023.

[32] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, *Inf. Sci.*, vol. 619, pp. 679–694, 2023.

[33] Y. Xu, H. Su, G. Ma, and X. Liu, A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context, *Complex Intell. Syst.*, vol. 9, no. 1, pp. 951–963, 2023.

Yan Zhao received the BS degree in electronic information engineering and the master degree in electronic information from Linyi University, Linyi, China, in 2019 and 2023, respectively. His current research interests include pattern recognition, multi-sensor information fusion, and digital signal processing.

Ming Guo received the BS degree in applied mathematics from Linyi University, Linyi, China, in 2010, and the MS degree in systems theory from Northeastern University, Shenyang, China, in 2012, and the PhD degree in control theory and control engineering from Dalian University of Technology, Dalian, China, in 2018. He is currently an associate professor at the School of Automation and Electrical Engineering, Linyi University, Linyi, China. His research interests include pattern recognition, body sensor networks, and bioinformatics.

Xiangyong Chen received the BS degree in applied mathematics from Linyi University, Linyi, China, in 2006, and the master and PhD degrees in control theory and control engineering from Northeastern University, Shenyang, China, in 2008 and 2012, respectively. From 2017 to 2018, he was a visiting scholar at the Department of Electrical Engineering, Yeungnam University, Gyeongsan, Republic of Korea. From 2014 to 2019, he was a postdoctoral fellow at the School of Mathematics, Southeast University, Nanjing, China. He is a professor at the School of Automation and Electrical Engineering, Linyi University, Linyi, China. His current research interests include complex dynamic system and complex networks, differential game and optimal control, and synchronization control of chaotic systems.

Jianqiang Sun received the PhD degree in applied mathematics from University of Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an associate professor at the School of Automation and Electrical Engineering, Linyi University, Linyi, China. His current research interests include bioinformatics, data mining, and deep learning.

Jianlong Qiu received the PhD degree in applied mathematics from Southeast University, Nanjing, China, in 2007. He is currently a professor at the School of Automation and Electrical Engineering, Linyi University, Linyi, China. He was a visiting scholar at Stevens Institute of Technology, Hoboken, NJ, USA, and University of Rhode Island, Kingston, RI, USA, where he was involved in collaborative research. His current research interests involve complex networks, stability theory, neural network, genetic networks, nonlinear system, and applied mathematics.