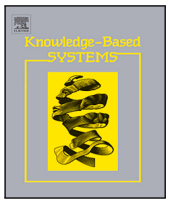
Knowledge-Based Systems 308 (2025) 112744

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Affective body expression recognition framework based on temporal and spatial fusion features

Tao Wang [a,1], Shuang Liu [a,*,1], Feng He [a,b], Minghao Du [a], Weina Dai [b], Yufeng Ke [a], Dong Ming [a,b]

[a] Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China
[b] Department of Biomedical Engineering, College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Keywords:
Emotion recognition
Body movement
Temporal–spatial feature
Energy model
Feature fusion

ABSTRACT

Affective body expression recognition technology enables machines to interpret non-verbal emotional signals from human movements, which is crucial for facilitating natural and empathetic human–machine interaction (HCI). This work proposes a new framework for emotion recognition from body movements, providing a universal and effective solution for decoding the temporal–spatial mapping between emotions and body expressions. Compared with previous studies, our approach extracted interpretable temporal and spatial features by constructing a body expression energy model (BEEM) and a multi-input symmetric positive definite matrix network (MSPDnet). In particular, the temporal features extracted from the BEEM reveal the energy distribution, dynamical complexity, and frequency activity of the body expression under different emotions, while the spatial features obtained by MSPDnet capture the spatial Riemannian properties between body joints. Furthermore, this paper introduces an attentional temporal–spatial feature fusion (ATSFF) algorithm to adaptively fuse temporal and spatial features with different semantics and scales, significantly improving the discriminability and generalizability of the fused features. The proposed method achieves recognition accuracies over 90% across four public datasets, outperforming most state-of-the-art approaches.

1. Introduction

Emotion recognition plays a crucial role in human survival and social interaction. With the development of human–machine interaction (HCI) technology, the timely and effective understanding of users' emotions has become an important factor in improving HCI efficiency and user experience [1]. Therefore, automatic emotion recognition has received extensive attention in recent years.

Although significant progress has been made in emotion recognition using different approaches, including text [2,3], vocal expressions [4], facial expressions [5], and electroencephalogram (EEG) [6], there have been few studies on emotion recognition from body expressions. Body expression can be understood as the movement of extremities, the torso, and various other human body parts, which is one of the most important marks of the human cognitive state [7]. It has been found that 65% of human emotional expressions are influenced by non-verbal signals, such as body postures and motions [8]. Moreover, body movements have been proven to provide comparable recognition accuracy relative to facial expression [9,10], and there is a special mapping between human movements and emotions [11]. With the advancement of motion capture technology, it is now possible to precisely track the trajectories of major human joints in a three-dimensional (3D) space [12]. This has further promoted research on emotion recognition based on 3D body skeletal data.

Although enormous efforts have been made in affective body expression recognition, the current solutions still face the following issues. First, for the temporal and spatial analysis of body expressions, previous studies have often extracted the kinematic features of body movements, such as speed, arm swing, and head movement, as affective temporal–spatial representations [13–15]. However, these features were concentrated on the certain-frame data of a few joints, and thus only shallow features could be extracted. Many subsequent studies have instead utilized deep learning models to extract emotional temporal and spatial information in body expressions. For example, Bhatia et al. [16] used a hybrid architecture combining LSTM and MLP to classify four emotions based on affective body expressions. Karumuri et al. used a multi-input CNN structure for emotion recognition from body skeleton data encoded by images [17,18]. Bhattacharya et al. [19] improved the Spatial Temporal Graph Convolutional Network (ST-GCN) model

* Corresponding author.
E-mail addresses: taowang2021@tju.edu.cn (T. Wang), shuangliu@tju.edu.cn (S. Liu), heaven@tju.edu.cn (F. He), minghaodu@tju.edu.cn (M. Du), wndai@tju.edu.cn (W. Dai), clarenceke@tju.edu.cn (Y. Ke), richardming@tju.edu.cn (D. Ming).
[1] Tao Wang and Shuang Liu contributed equally to this work.

https://doi.org/10.1016/j.knosys.2024.112744
Received 8 April 2024; Received in revised form 26 October 2024; Accepted 12 November 2024
Available online 20 November 2024

proposed in [20], and used the model to learn temporal–spatial patterns from skeletal data of body expressions. However, such methods often rely on deep neural networks (DNNs) with a large number of parameters, leading to overfitting of the model and reduced interpretability. Therefore, there is a need to further propose more interpretable models and features to understand and analyze the complex mapping relationship between body expressions and emotional states. Second, existing methods for affective body expression recognition often rely on temporal–spatial fusion features. However, these approaches typically perform feature fusion using simple operations such as concatenation or addition. This results in fixed linear mappings of the fused features, thereby neglecting the diverse semantic and scale information inherent in temporal and spatial features. Therefore, these methods fail to efficiently and dynamically integrate crucial temporal–spatial features.

To address the existing problem, this work proposes an innovative framework for affective body expression recognition. The framework extracts advanced and interpretable temporal–spatial features, providing a general and effective solution for decoding the complex mapping between emotions and body expressions. Considering the first question, for the temporal feature analysis, we construct an interpretable Body Expression Energy Model (BEEM) to quantify the effects of different emotions on the energy distribution, dynamic complexity and frequency activity of posture movements. For the spatial feature analysis, we encode spatial correlations between body joints using position and angle covariance matrices. Meanwhile, we propose the Multi-Input Symmetric Positive Definite Matrix Network (MSPDnet) to learns the spatial Riemannian properties from the position and angle matrices. To address the second problem, namely fusing temporal and spatial features with different semantics and scales, and mitigating the overfitting problem in feature fusion, we introduce the Attentional Temporal–Spatial Feature Fusion (ATSFF) algorithm. The algorithm dynamically extracts contextual information across global and local scales, enhancing the discriminability and generalizability of the fused temporal–spatial features.

We evaluated the performance and generalization capability of our method on four public datasets, which were collected using different devices (e.g. Kinect and Mocap). On all datasets, our method achieved an emotion recognition rate of over 90%, surpassing the state-of-the-art approaches. The primary contributions of this paper can be summarized as follows:

- An interpretable BEEM is constructed to quantify dynamic energy evolution during body movements. The optimal temporal feature patterns extracted from the BEEM decode the energy distribution, dynamic complexity, and frequency activity properties of body expressions under different emotions.
- The MSPDnet consisting of two parallel shallow matrix networks is proposed to learn the spatial Riemannian properties between body joints. It not only jointly processes input multi-dimensional posture covariance matrices, but also maintains the spatial properties encoded in the matrix structure during the mapping process of the network.
- The ATSFF algorithm is proposed to fuse temporal and spatial features with inconsistent semantics and scales, while enabling the model to adaptively extract fused features at both global and local scales, significantly improving the performance of feature fusion.

The rest of this paper is organized as follows. Section 2 reviews related research on affective body expression recognition. Section 3 describes the proposed framework in detail, including temporal and spatial feature extraction and the feature fusion optimization algorithm. The experimental results of our approach are presented in Section 4. Finally, conclusions and future work are discussed in Section 5.

2. Related work

2.1. Affective body expression recognition

Affective body expression recognition is the process of interpreting and understanding human emotions through the analysis of body postures and movements. In recent years, researchers have conducted in-depth studies on the relationship between body expressions and emotional states, and skeletal data analysis in particular has received widespread attention for its intuitive and effective approach [21,22]. With the development of affordable and portable depth sensors and wearable devices, researchers can more easily extract skeletal information of body movements, especially using devices like Kinect and Motion Capture (MoCap). The Kinect is a depth-sensing input device that uses an infrared projector and camera to create a depth map of objects by measuring the reflection time of light points, thus capturing the 3D skeletal data of human motions [23]. The MoCap is a generic term for motion capture system. The system employs wearable sensors, such as accelerometers, gyroscopes, and magnetometers, to capture 3D skeleton data during body movements [12].

Because of the significant advantages of Kinect and MoCap in terms of extraction accuracy, real-time performance and the ability to capture complex movements, more researchers have used these technologies to extract skeletal information of body expressions and integrate it with sophisticated machine learning methods for affective body expression recognition. For instance, Ahmed et al. [24] used Kinect to extract skeletal information of body movements, and the authors used statistical and genetic algorithms to achieve effective two-layer feature extraction and combined various machine learning algorithms, such as Decision Tree (DT) and Gaussian Plain Bayes (GNB), for the recognition of five emotions. Zhang et al. [25] collected seven types of emotional body expressions using Kinect and proposed a stack LSTM network based on attention (AS-LSTM) for emotion recognition from body movements. In another study, Daoudi et al. [26] used MoCap to perform Riemannian centroid calculations of raw joint trajectories. By utilizing the log-Euclidean Riemannian metric between the test data and class centers, the authors classified five emotions using a nearest-neighbor classifier. Bhattacharya et al. [27] introduced a semi-supervised learning approach for classifying human-perceived emotions from walking styles captured by the MoCap system.

These studies highlight the importance of leveraging advanced sensor technologies like MoCap and Kinect for emotional expression analysis. Therefore, this work proposes a new method for recognizing affective body expressions based on skeletal data using multiple datasets from MoCap and Kinect, aiming to provide a general and effective solution for decoding the mapping relationship between emotions and body expressions.

2.2. Temporal and spatial analysis of affective body expression

Many previous studies have used deep learning models to extract temporal–spatial features of body expressions. These models often extract features from affective body expressions through complex network mappings. For example, Ghaleb et al. [28] encoded the body joints into a graphical format, which was then processed with Graph Convolutional Networks (GCNs) to obtain temporal–spatial features of the body expressions. Sapinski et al. [29] tested the performance of various neural network architectures for affective body expression recognition, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-term Memory (LSTM). The results of the study show the superior performance of the RNN-LSTM model. Avola et al. [30] constructed a multi-branch architecture that employs LSTM to analyze temporal features and uses MLP to extract global features. Zacharatos et al. [31] encoded 3D skeleton motion data into 2D images and used the pre-trained convolutional neural network Inception V3 to classify happy and sad emotions. However, these approaches

rely on deep neural networks with extensive model parameters, which can lead to overfitting and lack of interpretability. Therefore, more and more researchers focus on proposing interpretable models and features to understand and analyze the relationship between body expressions and emotional states. For example, Oguz et al. [32] proposed a method based on sliding windows to extract temporal–spatial features of body expressions. The authors segmented the body skeleton sequence into different lengths of windows and then extracted features such as mean, root mean square, continuous wavelet transform, and joint neighborhood distance from each window. Using these methods, the authors extracted interpretable spatiotemporal features for affective body expression recognition. Some researchers have focussed on using the energy generated by body movements to investigate body expressions. For instance, Li et al. [33] obtained motion energy from a 3D body skeleton sequence by calculating coordinate differences. Then, they performed a discrete Fourier transform on the energy model to extract time–frequency features used to identify three emotions. In another study, Kacem et al. [34] used covariance matrices to encode 3D body movements, and then they proposed a geometric-perceptual (dis-) similarity metric that extracts spatial features from the covariance matrices for emotion recognition. Given that, this study proposes the BEEM and MSPDnet based on matrix network architecture to extract effective and interpretable temporal and spatial features for affective body expression recognition. Furthermore, we introduce a novel feature fusion algorithm ATSFF to adaptively fuse temporal–spatial features. The proposed method provides a generic and effective solution for decoding the complex mapping between emotions and body expressions.

3. Method

3.1. Method overview

The overview of the proposed framework is given in Fig. 1. It comprises three main components: temporal feature extraction, spatial feature extraction, and the fusion of temporal and spatial features for emotion recognition. For the temporal feature extraction (Section 3.2), an interpretable energy model BEEM is constructed to quantify the temporal dynamic energy evolution of the body movement across time (Section 3.2.1), which is illustrated in Fig. 1(a). Then, multiple types of features extracted from the BEEM were systematically evaluated to determine the impact of different emotions on energy distribution, dynamic complexity, and frequency activity of body movements (Section 3.2.2), which is shown in Fig. 1(b). For the spatial feature extraction (Section 3.3), we first encode the spatial Riemannian correlations between different joints using position and angle covariance matrices (Section 3.3.1). Subsequently, the MSPDnet is proposed to learn the spatial body expressions encoded in the multi-dimensional posture covariance matrices (Section 3.3.2). The above processes are shown in Fig. 1(c) and (d). To fuse the above obtained temporal and spatial features with different semantics and scales, we introduce the ATSFF algorithm (Section 3.4). The algorithm can dynamically extract the contextual information of the fused features at global and local scales. The fused temporal–spatial features are used for emotion recognition from body expressions, as shown in Fig. 1(e). Table 1 provides detailed descriptions of the main symbols.

3.2. Temporal feature extraction

3.2.1. Body expression energy model (BEEM)

In temporal analysis, the energy generated by body expression is considered an important measure of emotional state [35]. Therefore, this study hypothesize that an energy model can quantify the correlation between body expressions and emotions, thereby constructing an interpretative model for the temporal analysis of body expressions. For this purpose, we adopted the theoretical framework of mechanical energy and developed a body expression energy model (BEEM).

We define the BEEM of the ith joint in the fth frame during body expression as follows:

$$E_i^f = E_{k,i}^f + E_{p,i}^f = \frac{1}{2} m \left(\mathbf{V}_i^f \right)^2 + mg\mathbf{H}_i^f \tag{1}$$

where $i \in [1, N]$ and $f \in [1, F]$, and N and F denote the total numbers of joints and frames respectively. $E_{k,i}^f$ denotes the kinetic energy, and $E_{p,i}^f$ denotes the potential energy. $\mathbf{V}_i^f$ represents the velocity of each joint, while $\mathbf{H}_i^f$ represents the Euclidean distance between the position of each joint and its corresponding neutral posture. In this study, we assume that all body joints have the same mass m and ignore the gravitational acceleration g. Therefore, the BEEM is composed of the kinetic energy curve $\mathbf{V}$ and the potential energy curve $\mathbf{H}$.

First, we construct the kinetic energy curve $\mathbf{V}$. Body movement is approximated as the evolution of N joints in 3D space (x-, y-, and z-axis) over time. Furthermore, angular velocity has been shown to be an effective representation in affective body expression [36]. Therefore, we combine the displacement velocity and angular velocity of joints by calculating the finite differences between consecutive frames (i.e., between frame f and frame $f-1$). We assume that all joints have zero velocity at the initial frame $f = 0$. The result is defined as a kinetic energy curve, which describes the kinetic energy variation of body expressions over time. The kinetic energy curve is defined as follows:

$$\mathbf{V} = \begin{pmatrix} P_1^2 - P_1^1 & \cdots & P_N^2 - P_N^1 \\ \vdots & \ddots & \vdots \\ P_1^F - P_1^{F-1} & \cdots & P_N^F - P_N^{F-1} \end{pmatrix}, \quad P \in \mathbb{R}^6 \tag{2}$$

where $P_i^f = [x, y, z, \delta, \theta, \varphi]$ indicates the position and angle parameter of the ith joint in the fth frame.

To investigate the potential energy during body movements, this study initially defines a neutral posture as the baseline for measuring potential energy changes. The neutral posture is identified as a relaxed standing position with the head raised, both arms naturally hanging down by the thighs, and both legs remain upright. For this purpose, we define the following six vectors:

$$\vec{l}_i = \begin{cases} \text{head} - \text{shoulder}_{\text{center}} & \text{for } i = 1 \\ \text{shoulder}_{\text{center}} - \text{hip}_{\text{center}} & \text{for } i = 2 \\ \text{shoulder}_j - \text{wrist}_j & \text{for } i = 3, 4,\ j\text{:left or right} \\ \text{hip}_k - \text{ankle}_k & \text{for } i = 5, 6,\ k\text{:left or right} \end{cases} \tag{3}$$

Subsequently, in all movement segments of each subject, we obtained the individual's neutral posture $P_{neutral}$ using the following definition:

$$\alpha_{sum} = \sum_{i=1}^{6} \arccos \left(\frac{\vec{l}_i \cdot \vec{n}}{\|\vec{l}_i\| \cdot \|\vec{n}\|} \right) \tag{4}$$

$$S_{min} = \underset{f \in F, |S_{min}| = t}{\operatorname{argmin}} (\alpha_{sum}^f) \tag{5}$$

$$P_{neutral,i} = \frac{1}{|S_{\min}|} \sum_{f \in S_{\min}} P_i^f \tag{6}$$

where $\vec{n}$ is the normal to the horizontal plane. F is the set of all posture frame sequences for each individual, and $S_{\min}$ is the set of frame data that contains the frame sequence corresponding to the smallest t values of α_{sum}, and t is taken to be 6 in this paper. P_i^f is the position and angle parameter of the ith joint in the fth frame.

We obtained the neutral posture corresponding to each individual by the above equation. Finally, for each body expression segment, we integrate the 3D positional and angular information of each joint and compute their Euclidean distances from the neutral posture in each frame. This result is known as the potential energy curve, which

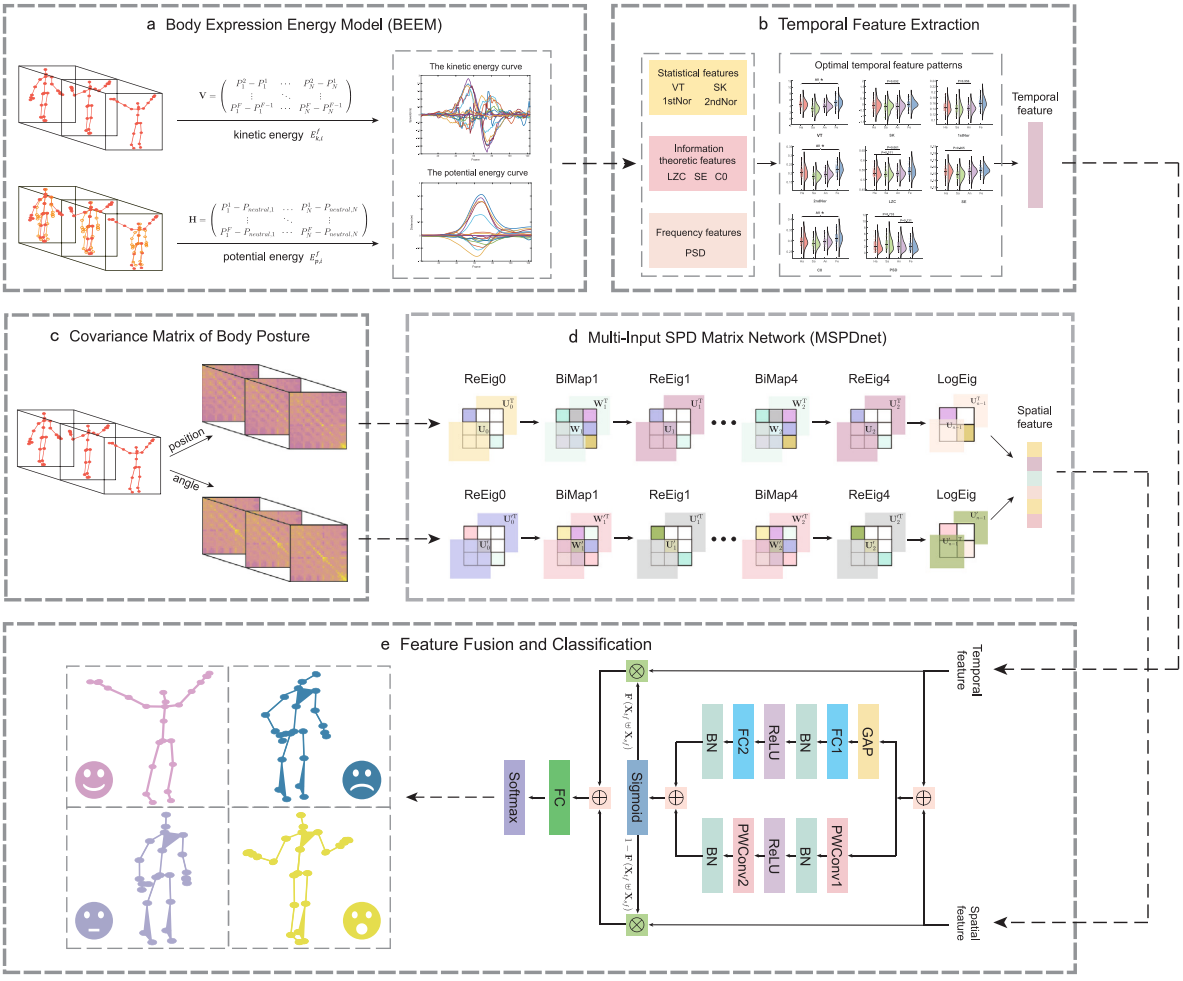


Fig. 1. The framework of the proposed method. In the extraction of temporal features, **(a)** the BEEM is established to quantify the dynamic energy evolution by combining the curves of kinetic energy and potential energy; **(b)** multiple types of features are then utilized to obtain the optimal temporal feature patterns from the BEEM. VT (variance trend), SK (skewness), 1stNor (first differences), 2ndNor (second differences), LZC (Lempel–Ziv complexity), SE (Shannon entropy), C0 (C0-complexity), PSD (power spectral density). In the extraction of spatial features, **(c)** the position and angle information of the posture skeleton sequences are first encoded into covariance matrices with different dimensions; **(d)** the MSPDnet is then proposed to map multi-dimensional covariance matrices onto the Riemannian manifold and jointly extract the spatial features. Finally, **(e)** the temporal and spatial features are fused by the ATSFF, and then fed into the fully connected (FC) and softmax layer for emotion recognition.

Table 1
Notations and descriptions.

Symbol	Description	Symbol	Description
E_i^f	Energy of ith joint in fth frame	$E_{k,i}^f$	Kinetic energy of ith joint in fth frame
$E_{p,i}^f$	Potential energy of ith joint in fth frame	V_i^f	Velocity of ith joint in fth frame
H_i^f	Euclidean distance between ith joint and its neutral posture	P_i^f	Position and angle parameter of ith joint in fth frame
$P_{neutral}$	Neutral posture	VT	Variance Trend
SK	Skewness	$1stNor$	First normalized difference
$2ndNor$	Second normalized difference	LZC	Lempel–Ziv complexity
SE	Shannon entropy	$C0$	C0-complexity
PSD	Power spectral density	C	Covariance matrix of the skeletal segment
X_{tf}	Temporal feature extracted from BEEM	X_{sf}	Spatial feature extracted from MSPDnet
$G(X)$	Global feature context	$L(X)$	Local channel context
TS	Fused temporal and spatial feature	X_s	Statistical features
X_e	Information-theoretic features	X_f	Frequency features
C_p	Covariance matrix of 3D position of body expression	C_a	Covariance matrix of 3D angle of body expression

effectively reflects the displacement $\mathbf{H}$ of each joint relative to the neutral posture, and is defined as follows:

$$\mathbf{H} = \begin{pmatrix} P_1^1 - P_{neutral,1} & \cdots & P_N^1 - P_{neutral,N} \\ \vdots & \ddots & \vdots \\ P_1^F - P_{neutral,1} & \cdots & P_N^F - P_{neutral,N} \end{pmatrix}, P \in \mathbb{R}^6 \tag{7}$$

We combine the kinetic and potential energy curves to obtain the BEEM, which is shown in Fig. 1(a). This model quantifies the dynamic energy evolution in affective body expressions and provides us with an interpretable model for decoding the temporal mapping between emotions in body expressions.

3.2.2. Temporal feature analysis

In this section, we extract multiple types of features from BEEM to form optimal temporal feature patterns for interpreting and analyzing affective body expressions. Given the multifaceted and complex nature of emotional states, we extracted various features, including statistical, information-theoretic, and frequency features. We used a 2-second sliding window with a 1.5-second overlap to segment the original skeletal sequence into multiple segments and extracted these features from each segment. Based on these features, we examined how emotional states are expressed through changes in energy distribution, dynamic complexity, and frequency activity of body movements.

(1) Statistical features have often been used in the analysis of human postures [37,38]. For each dimension sequence $\mathbf{e}(j) \in \mathbb{R}^f$ in the BEEM, the following four statistical features are extracted, where f denotes the sequence length.

- Variance Trend (VT): it measures the stability of data fluctuations over consecutive time windows. We divide a posture segment into several small windows. The variance $\delta^2(\mathbf{e})$ for each window is calculated, and then the absolute value of the difference between the variance of each window and its neighboring window is summed up. This result describes the variance trend over the duration of body expression, which is computed as follows:

$$\delta^2(\mathbf{e}) = \frac{1}{f-1}\sum_{j=1}^{f}\left(\mathbf{e}_j - \mu_{\mathbf{e}}\right)^2 \tag{8}$$

$$VT = \sum_{j=2}^{f/m}\left(\left|\delta_j^2(\mathbf{e}) - \delta_{j-1}^2(\mathbf{e})\right|\right) \tag{9}$$

  where $\mu_{\mathbf{e}}$ is the mean of the signal, m is the number of windows into which the posture segment is divided, which is set to 10. A high VT indicates large fluctuations in the energy distribution of body expressions, which may suggest unstable emotional states or intense body motions.
- Skewness (SK): it describes the asymmetry of energy fluctuations generated by affective body expressions. Positive skewness indicates that high-energy states are dominant, suggesting high arousal emotions. On the other hand, negative skewness may be associated with low arousal emotional states.

$$SK = \frac{\frac{1}{f}\sum_{j=1}^{f}\left(\mathbf{e}_j - \mu_{\mathbf{e}}\right)^3}{\left\{\frac{1}{f}\sum_{j=1}^{f}\left(\mathbf{e}_j - \mu_{\mathbf{e}}\right)^2\right\}^{3/2}} \tag{10}$$

- The first and second normalized differences (1stNor, 2ndNor): The 1stNor measures the immediate change in body motions. The 2ndNor quantifies the rate of change in posture movements. These indicators can detect sudden emotional responses or gradual emotional shifts. The 1stNor can be calculated by:

$$\mathbf{e}' = \frac{\mathbf{e}(j) - \mu_{\mathbf{e}}}{\delta(\mathbf{e})} \tag{11}$$

$$1stNor = \frac{1}{f-1}\sum_{j=1}^{f-1} \mid \mathbf{e}'(j+1) - \mathbf{e}'(j) \mid \tag{12}$$

- The 2ndNor can be calculated by:

$$2nd\,Nor = \frac{1}{f-2}\sum_{j=1}^{f-2} \mid \mathbf{e}'(j+2) - \mathbf{e}'(j) \mid \tag{13}$$

(2) The three following information-theoretic features are extracted from the BEEM.

- Lempel–Ziv Complexity (LZC): It indicates the richness and unpredictability of the posture energy signals. Higher values indicate a high complexity of body expression. First, for the data of each dimension in the BEEM, we use the average value as the threshold to binarize it and the length of the processed sequence is n. Next, the binarized sequence is decomposed into q blocks, which are used to compute the LZC as follows:

$$LZC = \frac{q\log_2 n}{n} \tag{14}$$

- Shannon entropy (SE): It quantifies the randomness of body expressions. High entropy values indicate the diversity and high complexity of body motions.

$$SE = -\sum_{i=1}^{K} p(i)\log(p(i)) \tag{15}$$

  where K represents the number of unique values in each dimension of the BEEM and $p(i)$ is the corresponding probability for these unique values.
- C0-complexity: It measures the irregularity of affective body expressions. For each dimension sequence $\mathbf{e}(j) \in \mathbb{R}^f$ in the BEEM, the mean amplitude of the power spectrum of $\mathbf{e}(j)$ can be obtained as follows:

$$D = \frac{1}{f}\sum_{k=0}^{f-1}|\mathbf{e}(k)|^2 \tag{16}$$

  where $\mathbf{e}(k)$ is the fast Fourier transform (FFT) of $\mathbf{e}(j)$. A new spectrum is constructed using $\mathbf{e}(k)$ and D as follows:

$$\mathbf{y}(k) = \begin{cases} \mathbf{e}(k) & |\mathbf{e}(k)|^2 > D \\ 0 & |\mathbf{e}(k)|^2 \le D \end{cases} \tag{17}$$

  The C0-complexity of $\mathbf{e}(j)$ can be calculated by:

$$C0 = \frac{A_1}{A_0} = \frac{\sum_{j=0}^{f-1}|\mathbf{e}(j) - \mathbf{y}(j)|^2}{\sum_{j=0}^{f-1}|\mathbf{e}(j)|^2} \tag{18}$$

  where $\mathbf{y}(j)$ is the inverse Fourier transform of $\mathbf{y}(k)$, and A_1 and A_0 are the powers of irregular and regular parts of $\mathbf{e}(j)$, respectively.

(3) The absolute power spectral density (PSD) of the low frequency (0.1–3 Hz) and high-frequency (3–10 Hz) was calculated using Welch's FFT technique [39]. The reason we employ both low-frequency and high-frequency analysis is that human motion can be divided into micro-movements (rapid and subtle movements) and macro-movements (slow and wide-range movements). Therefore, low-frequency PSD captures macro-movements with slow changes in affective body expression, while high-frequency PSD captures the quicker micro-movements. A data window is used in each dimension sequence of BEEM.

Let $xd(t)$ be the sequence, where $d = 1, 2, 3 \ldots L$ (signal intervals) and M is interval length. Hence, the definition of power spectral density utilizing the Welch method is as follows:

$$d(f) = \frac{1}{MU}\left|\sum_{t=0}^{M-1} xd(t)w(t)e^{-j2\pi f}\right|^2 \tag{19}$$

which U stands for normalization factor for power in window function

$$U = \frac{1}{M}\sum_{t=0}^{M-1}|w(t)|^2 \tag{20}$$

where $w(t)$ represents windowed data, the Welch power spectrum is calculated as the average over modified periodograms defined by:

$$Welch(f) = \frac{1}{L}\sum_{i=0}^{L-1} d(f) \tag{21}$$

In this study, each posture segment was divided into 10 segments without overlap. Each segment was subjected to the Hamming window, and the resulting periodograms were averaged to determine the PSD estimate. The spectra are calculated by the periodogram method using a 128-point FFT and periodic Hamming windows.

Subsequently, we will use the above three types of features and their fused features to recognize affective body expression, aiming to construct the optimal temporal feature patterns. To obtain the more generalized optimal temporal feature patterns, we conducted extensive experiments on six common classifiers. We evaluated K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP) classifiers across four datasets. Each classifier was tested for performance on statistical, information-theoretic, frequency features, and their fusion features. We also combine an infinite feature selection algorithm [40] for dimensionality reduction. To prevent overlooking important features with small or negative values during classification, the min–max normalization method was used to normalize the features.

3.3. Spatial feature extraction

3.3.1. Posture covariance matrix

In the above temporal features extraction, each joint is studied independently, ignoring the significance of the spatial correlations between joints for body expression recognition. Therefore, in this section, we introduce the covariance matrix to capture the spatial correlation between different joints.

Assume $\mathbf{x} \in \mathbb{R}^d$ denotes a d-dimensional feature vector of the 3D position information of the joints, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_f] \in \mathbb{R}^{d\times f}$ denotes a set containing the d-dimensional feature vectors of f frame skeletal sequences. The covariance matrix of the skeletal segment $\mathbf{X}$ is defined by:

$$\mathbf{C} = \frac{1}{f-1}\sum_{j=1}^{f} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^{\mathrm{T}} \tag{22}$$

where $\boldsymbol{\mu}$ is the mean of $\mathbf{x}_f$. In this study, considering the importance of angle information of each joint for body expression recognition, we encoded the 3D position and angle information of the skeleton sequence separately using the covariance matrix. In addition, we used a 2 s sliding window with an overlap of 1.5 s to divide the original skeleton sequence into segments. In each posture segment, we extracted the position and angle covariance matrices.

3.3.2. Multi-input SPD matrix network

A non-singular covariance matrix belongs to the SPD matrices, which form a connected Riemannian manifold Sym_d^+ [41]. Previous research has shown that Riemannian neural network [42] can extract more discriminative spatial features from the SPD matrix, significantly improving classification performance. Therefore, based on the Riemannian network architecture [42], we improve the arrangement of the network layers and construct a two-branch network structure, thus proposing the multi-input symmetric positive definite matrix network (MSPDnet). The MSPDnet can jointly process the multi-dimensional covariance descriptors encoded with position and angle information and the network architecture of MSPDnet is detailed in Fig. 1(d). Each branch network is composed of multiple parallel eigenvalue rectification (ReEig) layers, bilinear mapping (BiMap) layers, and eigenvalue logarithm (LogEig) layers. The inputs are covariance matrices of 3D position and angle information. At the network output, the mapped covariance matrices are vectorized and joined to form a one-dimensional vector. It is then fused with the temporal feature by the ATSFF algorithm.

(1) ReEig layer: The input covariance matrix calculated by (22) may belong to the Symmetric Positive Semi-Definite (SPSD) matrices. Therefore, the first layer of the network is ReEig, which serves to normalize the covariance matrix. Furthermore, a ReEig layer follows each BiMap layer to ensure the mapped matrix remains on the Riemannian manifold and to introduce a nonlinear mapping process. The ReEig layer is defined as follows:

$$\mathbf{C}_{r,n} = f_r(\mathbf{C}_{n-1}, \varepsilon) = \mathbf{U}_{n-1}\ \mathrm{Max}(\varepsilon\mathbf{I}, \Lambda_{n-1})\,\mathbf{U}_{n-1}^{\mathrm{T}} \tag{23}$$

$$\mathrm{Max}(\varepsilon\mathbf{I}, \Lambda_{n-1}) = \mathbf{E}(i,i) = \begin{cases} \Lambda(i,i), & \text{if } \Lambda(i,i) > \varepsilon \\ \varepsilon, & \text{if } \Lambda(i,i) \le \varepsilon \end{cases} \tag{24}$$

where $\mathbf{U}_{n-1}$ and Λ_{n-1} denote the eigenvectors and eigenvalues of the input matrix $\mathbf{C}_{n-1}$ in nth layer respectively, and $\mathbf{I}$ is an identity matrix. The ε is a preset threshold which is used to replace null eigenvalues or small eigenvalues.

(2) BiMap layer: The function of the BiMap layer is to convert the SPD matrix into a matrix with higher discriminative properties, while preserving the geometric information encoded in the SPD matrix. The BiMap layer is defined as follows:

$$\mathbf{C}_{b,n} = f_b(\mathbf{C}_{n-1}, \mathbf{W}_n) = \mathbf{W}_n\mathbf{C}_{n-1}\mathbf{W}_n^{\mathrm{T}} \tag{25}$$

where $\mathbf{C}_{n-1} \in Sym_{d_{n-1}}^+$ is the input SPD matrix of size $d_{n-1}\times d_{n-1}$, and $\mathbf{W}_n \in \mathbb{R}^{d_n\times d_{n-1}}$ is the bilinear mapping matrix.

(3) LogEig layer: The LogEig layer imparts a Lie group structure to elements on Riemannian manifolds, allowing the matrix to be simplified into a flat space where traditional Euclidean computations can be applied [43]. The LogEig layer is defined as follows:

$$\mathbf{C}_{l,n} = f_l(\mathbf{C}_{n-1}) = \mathbf{Q}_{n-1}\log(\mathbf{A}_{n-1})\,\mathbf{Q}_{n-1}^{\mathrm{T}} \tag{26}$$

where $\mathbf{Q}_{n-1}$ and $\mathbf{A}_{n-1}$ denote the eigenvectors and eigenvalues of the input matrix respectively, and $\log(\cdot)$ is the matrix logarithm operation.

3.4. Fusion of temporal and spatial features

Considering the intricate correlation between body expressions and emotions, it is inadequate to focus solely on either temporal or spatial features. However, existing emotion recognition methods based on the fusion of temporal–spatial features often employ simplistic addition or concatenation. This approach merely offers a fixed linear aggregation of feature maps, ignoring the different semantic and scale information inherent in temporal and spatial features. To address this issue, inspired by the work of [44], we introduce an Attentional Temporal and Spatial Feature Fusion (ATSFF) algorithm. This algorithm dynamically fuses temporal and spatial features across different semantics and scales with attention mechanisms, significantly enhancing the performance of feature fusion.

Our algorithm uses a multiscale channel attention module to achieve attentional feature fusion across temporal and spatial dimensions. As illustrated in Fig. 2, the multiscale channel attention module innovatively designs a two-branch structure, which maintains network's sensitivity to global features while allowing for detailed processing of local features. Specifically, in the initial phase of the module, the $\oplus$ denotes the broadcasting addition of the temporal and spatial features. This initial integration provides a fused feature map, which ensures that subsequent attention mechanisms can consider both features with different semantics and scales simultaneously. In the left branch of the module, we introduce a combination of Global Average Pooling (GAP) and Fully Connected (FC) layers to extract global channel context of the fused temporal–spatial features, similar to SENet [45]. The GAP layer initially compresses the spatial information into global statistics for each channel, which are then further captured by the FC layer to grasp the global channel dependencies.

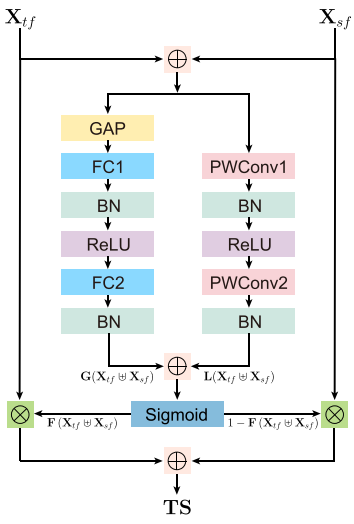


Fig. 2. The structure of the proposed multiscale channel attention module.

This process extracts global dependencies of the feature channels, allowing the model to understand broader, more generalized features. In the right branch of the module, we employ Point-wise Convolution (PWConv) [46] to extract local channel context. Through the 1×1 convolution mapping, we can capture the local channel dependencies of the fused features. This method captures fine-grained details and local dependencies within the feature channels. By combining global and local contexts, the model maintains sensitivity to both overarching patterns and detailed nuances. To maintain a lightweight design, we integrate the global context $\mathbf{G}(\mathbf{X})$ with the local context $\mathbf{L}(\mathbf{X})$ within the attention framework, employing a bottleneck structure for computation.

As shown in Fig. 2, given temporal and spatial feature X_{tf}, $X_{sf} \in \mathbb{R}^{C\times H\times W}$ with C channels, the fused feature can be expressed as follows:

$$\mathbf{TS} = \mathbf{F}(X_{tf} \uplus X_{sf}) \otimes X_{tf} + (1 - \mathbf{F}(X_{tf} \uplus X_{sf})) \otimes X_{sf} \tag{27}$$

where $\mathbf{TS} \in \mathbb{R}^{C\times H\times W}$ is the fused temporal and spatial feature, and $\uplus$ is the initial feature integration, which is computed based on broadcasting addition. The fusion weight of the temporal features generated by the attention module is $\mathbf{F}(X_{tf} \uplus X_{sf})$, which consists of real numbers between 0 and 1. The fusion weight $1 - \mathbf{F}(X_{tf} \uplus X_{sf})$ of the spatial features is also a real number. The fusion weights enable the network to perform weighted averaging between X_{tf} and X_{sf}. $\otimes$ denotes the element-wise multiplication after adjusting dimensions using padding, and $\mathbf{F}(X_{tf} \uplus X_{sf})$ is computed as follows:

$$\mathbf{F}(X_{tf} \uplus X_{sf}) = \sigma(\mathbf{G}(X_{tf} \uplus X_{sf}) \oplus \mathbf{L}(X_{tf} \uplus X_{sf})) \tag{28}$$

where σ is the Sigmoid function and $\oplus$ denotes the broadcasting addition. $\mathbf{G}(X_{tf} \uplus X_{sf}) \in \mathbb{R}^{C}$ is the global feature context and $\mathbf{L}(X_{tf} \uplus X_{sf}) \in \mathbb{R}^{C\times H\times W}$ is the local channel context, they can be computed as follows:

$$\mathbf{G}(X_{tf} \uplus X_{sf}) = \mathcal{B}\left(F_2\left(\delta\left(\mathcal{B}\left(F_1(g(X_{tf} \uplus X_{sf}))\right)\right)\right)\right) \tag{29}$$

$$\mathbf{L}(X_{tf} \uplus X_{sf}) = \mathcal{B}\left(\rho_2\left(\delta\left(\mathcal{B}\left(\rho_1(X_{tf} \uplus X_{sf})\right)\right)\right)\right) \tag{30}$$

where $g(X_{tf} \uplus X_{sf})$ is the GAP. F_1 and F_2 define FC1 and FC2. These two FC layers form the bottleneck structure, where F_1 is the dimension reduction layer, and F_2 is the dimension increasing layer. Similarly, ρ_1 and ρ_2 define PWConv1 and PWConv2. $\mathcal{B}$ denotes the Batch Normalization (BN) [47], and δ denotes the Rectified Linear Unit (ReLU) [48].

The temporal–spatial features fused by the multiscale channel attention module are fed into the FC and softmax layers to obtain the final prediction results, which is shown in Fig. 1(e). The optimization

Algorithm 1 Attentional temporal-spatial feature fusion

Input: temporal features: X_s, X_e, X_f
spatial features: C_p, C_a
learning rate: α_p, α_a, α_f
training set label: Y
Output: parameters of MSPDnet: ω_p, ω_a
parameters of FC layer: ω_f

initialization: iteration $t = 1$, $maxiter = 600$;
while $t \leq$ maxiter **do**
  Extract temporal features:
  # Combine statistical, information-theoretic, and frequency features extracted from the BEEM
  $X_{tf} = T(X_s,\ X_e,\ X_f)$
  Extract spatial features:
  # Use MSPDnet to process posture covariance descriptors
  $X_{sf}^t = S(C_p,\ C_a,\ \omega_p^t,\ \omega_a^t)$
  Fusing temporal and spatial features by Eq. (27)
  Compute output of the fully connected (FC) layer:
  # Feed the fused features into the FC layer for classification
  $f^t = FC\left(TS^t, \omega_f^t\right)$
  $Loss = L\left(f^t, Y\right)$
  Update the parameters of the MSPDnet:
  $\omega_p^{t+1} \leftarrow \omega_p^t - \alpha_p\left(\frac{\partial Loss}{\partial f^t} \cdot \frac{\partial f^t}{\partial X_{sf}^t} \cdot \frac{\partial X_{sf}^t}{\partial \omega_p^t}\right)$
  $\omega_a^{t+1} \leftarrow \omega_a^t - \alpha_a\left(\frac{\partial Loss}{\partial f^t} \cdot \frac{\partial f^t}{\partial X_{sf}^t} \cdot \frac{\partial X_{sf}^t}{\partial \omega_a^t}\right)$
  Update the parameters of the FC layer:
  $\nabla\omega_f^t = \frac{\partial Loss}{\partial f^t} \cdot \frac{\partial f^t}{\partial \omega_f^t}$
  $\omega_f^{t+1} \leftarrow \omega_f^t - \alpha_f \nabla\omega_f^t$
  $t = t + 1$
end while

scheme of the above temporal–spatial feature fusion algorithm ATSFF is shown in Algorithm 1. Where, X_s, X_e, and X_f represent the statistical, information-theoretic, and frequency features extracted from the BEEM, respectively. C_p and C_a are the covariance descriptors of 3D position and angle of body expression, respectively. α_p and α_a are the learning rates for processing the position and angle covariance matrices in the MSPDnet, respectively. α_f is the learning rate in the FC layer. In addition, ω_p and ω_a are the bilinear transformation matrices of two parallel networks in the MSPDnet. t denotes the number of iterations of the ATSFF algorithm and $maxiter$ is the maximum number of iterations. Further, the fused temporal–spatial feature vector is denoted as $TS^t = F(X_{tf} \uplus X_{sf}^t) \otimes X_{tf} + (1 - F(X_{tf} \uplus X_{sf}^t)) \otimes X_{sf}^t$, which is obtained from the multiscale channel attention module. X_{tf} is the optimal temporal feature pattern and X_{sf}^t is the spatial features extracted by MSPDnet. The mapping operation of the FC layer is represented by $FC\left(TS^t, \omega_f^t\right)$, where ω_f^t denotes a set of weight parameters of the FC layer. In addition, the loss function $L\left(f^t, Y\right)$ is defined as cross entropy and $\alpha_{(\cdot)}$ is the learning rate.

4. Results and discussion

In this section, we provide a detailed description of our experiments, mainly covering seven aspects: the introduction of the dataset (Section 4.1), the experimental settings (Section 4.2), the temporal feature analysis results (Section 4.3), the spatial feature analysis results (Section 4.4), the effectiveness evaluation of the attentional temporal–spatial feature fusion algorithm (Section 4.5), the performance assessment of the proposed framework (Section 4.6) and comparison with existing methods (Section 4.7).

4.1. Dataset

In this study, we aim to develop a universal and effective framework for affective body expression recognition. To ensure the effectiveness

Table 2
The description of the datasets used to evaluate the proposed method.

Datasets	Device	Subjects	Joints	Samples	Emotions
UCLIC	Mocap (Vicon MX)	13	32	183	Ha, Sa, An, Fe
EGBM	Kinect V2	16	25	560	Ha, Sa, Ne, An, Di, Fe, Su
KDAE	Mocap (Noitom PN)	22	72	1402	Ha, Sa, Ne, An, Di, Fe, Su
MPI	Mocap (Xsens MVN)	8	28	1447	An, Fe, Ha, Pr, Sa, Su, Re, Di, Ne, Am, Sh

Abbreviations for emotions: Am: Amusement, An: Anger, Di: Disgust, Fe: Fear, Ha: Happiness, Ne: Neutral, Pr: Pride, Re: Relief, Sa: Sadness, Sh: Shame, Su: Surprise.

and generalization of this framework, four publicly available datasets were used for validation and testing. These datasets include body expression data collected by various devices, such as Kinect and Mocap systems, from participants across diverse regions. Details of the four datasets used are listed in Table 2.

(1) UCLIC [49]: Captured with the Vicon MX Mocap system, this database has 183 emotional posture segments across 4 emotions, performed by 13 actors. It includes 3D position and rotation data for 32 body joints in each sample.

(2) EGBM [50]: Using the Kinect V2 sensor, this database contains 560 body motion samples for 7 emotions, acted by 16 Polish actors. Each sample includes 3D position and orientation data for 25 joints.

(3) KDAE [51]: Recorded with the Noitom Perception Neuron (PN) Mocap system, this dataset offers 1402 full-body expressions of 7 emotions, demonstrated by 22 Chinese actors. It tracks position and rotation for 72 anatomical nodes.

(4) MPI [52]: Using the Xsens MVN system, MPI contains 1,447 samples of 11 emotions, performed by 8 actors. Each segment has 3D data for 28 markers. Note: this database is highly imbalanced in terms of emotional expression.

4.2. Experimental settings

In the temporal feature analysis section (Section 3.2.2), to obtain the most generalized temporal feature patterns, we conducted extensive experiments using various types of temporal features and their fused features on six traditional classifiers. Table 3 lists the main parameter configurations used by the classifiers, which were optimized through grid search. In Section 3.3.2, we present the MSPDnet to extract spatial features, with four BiMap/ReEig blocks used for each branch of the network. In particular, the structure of each branch network is $\boldsymbol{C}_0 \rightarrow f_r \rightarrow f_b^{(1)} \rightarrow f_r^{(1)} \rightarrow f_b^{(2)} \rightarrow f_r^{(2)} \rightarrow f_b^{(3)} \rightarrow f_r^{(3)} \rightarrow f_b^{(4)} \rightarrow f_r^{(4)} \rightarrow f_l$, where f_r, f_b, f_l denote the ReEig, BiMap, and LogEig respectively. The transformation matrices $\mathbf{W}$ within the BiMap layers have been sized at $[d_{n-1} \times 50,\ 50 \times 40,\ 40 \times 30,\ 30 \times 20]$, respectively, with the d_{n-1} referring to the dimension of the input covariance matrix. The ReEig layer employs a rectification threshold ε that is set at 1e−8. The extracted optimal temporal feature patterns and spatial features are fused in the feature fusion algorithm ATSFF, which is shown in Fig. 1(e). In particular, the kernel sizes of FC1 and FC2 are $\frac{C}{r} \times C$ and $C \times \frac{C}{r}$, respectively. Similarly, the kernel sizes of PWConv1 and PWConv2 are $\frac{C}{r} \times C \times 1 \times 1$ and $C \times \frac{C}{r} \times 1 \times 1$, respectively. C is the channel number of the fused feature, and r is the channel reduction ratio and its value is 8 in this paper. The fused temporal–spatial features are fed into the FC and softmax layers to derive the final prediction. The structure employs an FC layer that has 128 hidden nodes, measuring its performance via the cross-entropy loss. The optimizer is Adam, with a learning rate $\alpha_{(\cdot)}$ set to 0.01. The training parameters specify a batch size of 64 and an epoch of 600, and the early stopping strategy is employed to mitigate overfitting.[2] The experimental process was conducted on two NVIDIA GeForce RTX 3090 GPUs, utilizing CUDA 12.1 through the TensorFlow API. All experiments utilize a 10-fold cross-validation approach to ensure robustness, with the classification performance being assessed through metrics such as accuracy, recall, F1-score, and AUC.

Table 3
The main parameters of the classifiers used in temporal feature analysis.

Algorithm	Parameter name and values
KNN	n_neighbors = 4, distance metric = 'Euclidean'
SVM	regularization: 0.2, gamma = 'scale', kernel = 'rbf'
RF	n_estimators = 200, criterion for split = 'Gini Impurity'
LDA	n_components = 1, solver = 'svd'
XGBoost	estimators: 1000, learning_rate = 0.1, tree_method = 'gpu_hist'
MLP	hidden_layer_sizes = [8, 16, 32], learning_rate = 0.01, max_iter = 100, activation = 'relu', solver = 'adam'

4.3. Results of temporal feature

In this section, we conduct extensive analyses of statistical, information-theoretic, and frequency features extracted from the BEEM (Section 4.3.1). Our aim is to reveal how emotional states are expressed through changes in the energy distribution, dynamic complexity, and frequency activity of body movements. In addition, we also employ multiple types of features and their fused features for affective body expression recognition, striving to construct optimal temporal feature patterns (Section 4.3.2).

4.3.1. Statistical analysis of multi-dimensional temporal feature

In this section, we thoroughly analyze the body expressions under various emotional states using multi-dimensional temporal features. The UCLIC dataset is used for visual analysis. For the four emotions (happiness, sadness, anger, surprise) in the dataset, we used one-way ANOVA for between-group comparisons, followed by post-hoc pairwise comparisons for features showing significant differences among the four emotion groups. These results are shown in Fig. 3. We perform Bonferroni correction to counteract the spurious positives caused by multiple comparisons [53].

As shown in Fig. 3, the VT of the BEEM was significantly different ($p < 0.05$) between groups for all four emotions. The highest VT was found for fear, followed by happiness, anger, and sadness. This indicates that body expressions may be more unstable and varied when experiencing fear. The SK results show higher skewness for fear and happiness, with no significant difference between sadness and anger ($p = 0.092$). This may indicate that fear and happiness might exhibit more extreme energy distribution shifts in affective posture expression. The 1stNor and 2ndNor results also indicate higher values for fear and happiness, suggesting rapid changes in body movement during these emotions.

In the information-theoretic features, the C0 shows the most significant differences across all four emotions. Additionally, the results for LZC, SE, and C0 indicate the highest complexity in posture movements under fear, suggesting complex and diversified patterns in fear-driven body expressions. In contrast, sadness exhibits the lowest complexity, reflecting slower and more limited body movements in sadness. Because the PSD only differed significantly between groups in the low-frequency band, only the results of PSD in the low-frequency band are shown. The PSD results revealed that in the low-frequency, the PSD values of sadness were significantly higher than those of the other emotions ($p < 0.05$), which suggests that the distribution of energy of

[2] Interested readers can contact the authors to obtain the code, and we will gladly provide the necessary resources.

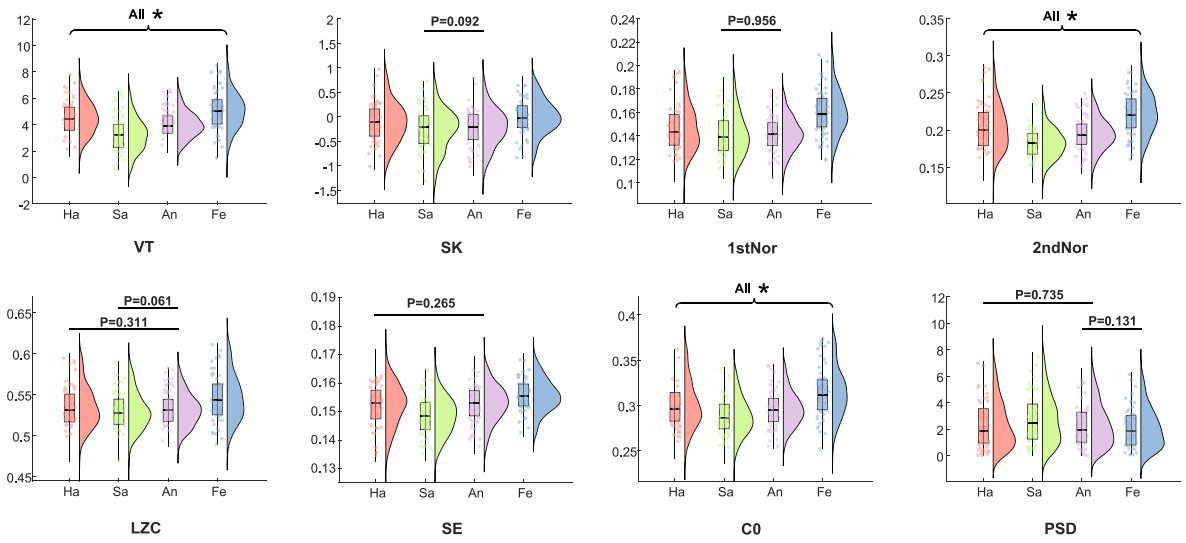


Fig. 3. The comparison of different features extracted from the BEEM on the UCLIC dataset over four emotions: Ha-happiness, Sa-sadness, An-anger, and Fe-fear. "All ∗" indicates significant differences ($p < 0.05$) in both between-group comparisons and post-hoc pairwise comparisons between each pair of groups. The horizontal lines with p-values indicate that the differences between the two groups are not significant in the post-hoc comparisons.

body postural movements in sadness tends to be in the low-frequency range. This may imply that body expressions under sadness have more macro-movements, i.e., slower and longer-lasting body motions.

Overall, the above results demonstrate three key conclusions: (1) The results of both statistical and information-theoretic features indicate that fear and happiness lead to highly unstable body expressions with wide movement ranges and rapid changes, showing high complexity. This may be due to the high arousal nature of these emotions, which are often accompanied by more pronounced physiological and behavioral responses, like escaping danger or engaging actively. This requires rapid and varied postural changes and more energy. (2) Sadness is typically associated with lower values in most features, indicating reduced intensity and complexity in body movements. This reduction may correspond to decreased physical activity and more restrained, introverted postures, aligning with findings from psychological studies on proximate psychological mechanisms [54]. (3) The low-frequency PSD analysis shows significantly higher values for sadness, which contrasts with the lower values observed in other features. This indicates that body expressions under sadness tend to be in the low-frequency range, suggesting that sadness-driven body expressions have slower and more sustained macro-movements.

4.3.2. Classification results of temporal features

In Section 4.3.1, we observed significant differences in various types of temporal features across different emotions, and there is strong complementarity between these features. Based on this finding, we further explore the performance of different features for affective body expression recognition and examined the enhancement achieved by fusing these features. Extensive experiments were conducted using six common classifiers to identify the optimal combination of temporal features.

Table 4 shows the performance of statistical, information-theoretic, frequency-domain features, and their fusion feature. The results indicate that the fused features generally provide higher accuracy than individual features. Specifically, the fused features achieved the best performance on all datasets with XGBoost and MLP classifiers. For example, XGBoost reached the highest accuracy of 89.02% and an AUC of 96.06% on the UCLIC dataset. For the EGBM, KDAE and MPI datasets, the MLP classifier obtained the highest accuracies of 85.29%, 83.74% and 72.30%, respectively. These results demonstrate the effectiveness of fused temporal features in improving emotion recognition performance. In summary, fusing temporal features significantly enhanced the recognition performance of affective body expressions. This fusion strategy effectively leveraged the strengths of individual features, improving the model's sensitivity to changes in emotional states.

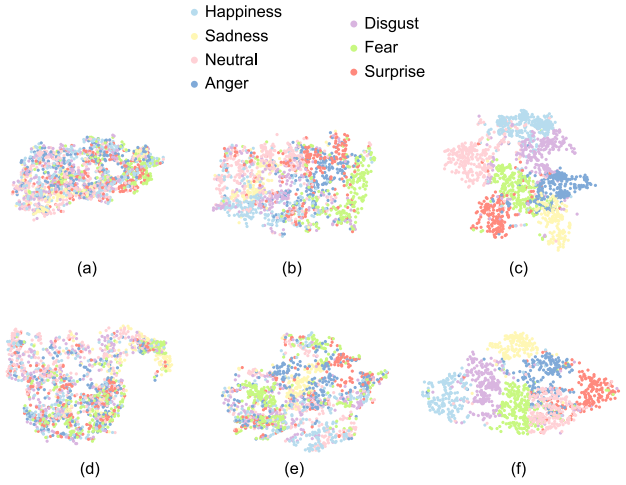


Fig. 4. The hierarchical visualization of the proposed MSPDnet on the KDAE dataset for the (a) input posture covariance matrix, (b) output of the ReEig2 layer, (c) output of the LogEig layer. (d)–(f) represent the results of the hierarchical visualization in the second branch network. Different colors indicate different emotional classes.

4.4. Results of spatial feature

To analyze the extracted spatial features after different mapping layer of the MSPDnet, we perform the hierarchical visualization of MSPDnet on the KDAE dataset using the t-distributed stochastic neighbor embedding (t-SNE) [55], which is shown in Fig. 4. The high-dimensional output matrices of different layers of MSPDnet are projected into the 2-D embedded space, and the emotional states are represented by different colors.

We first draw the raw position and angular covariance matrices in the network inputs, which are shown in Fig. 4(a) and (d). It can be observed that spatial features of different emotional classes are mixed haphazardly. Then, the input matrices are mapped through two BiMap/ReEig blocks, and the outputs of the ReEig2 layer are shown in Fig. 4(b) and (e). We can observe that the distribution of different

Table 4
The classification results of statistical, information-theoretic, frequency-domain, and temporal fusion features on various evaluation criteria (Accuracy, Recall, F1-score, and AUC). The best results are labeled in bold.

		UCLIC				EGBM				KDAE				MPI			
		Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC
Statistical Feature	KNN	79.68	78.68	78.10	83.68	75.20	74.40	74.34	83.04	72.12	71.83	69.60	75.56	56.81	54.96	54.36	62.92
	SVM	84.20	83.78	82.59	91.59	76.29	74.43	75.42	81.44	77.47	76.43	75.37	84.80	57.01	56.42	55.01	65.90
	RF	88.43	87.46	86.82	94.53	79.34	79.58	78.55	83.00	80.29	79.39	78.64	90.63	60.36	60.75	58.02	69.43
	LDA	79.23	78.30	77.46	89.04	72.21	69.79	69.31	77.45	74.39	73.67	72.25	83.25	54.34	53.34	51.98	61.10
	XGBoost	87.68	86.46	86.14	96.02	80.15	80.01	79.63	87.60	81.29	80.91	79.89	90.30	63.85	63.95	61.66	71.18
	MLP	86.15	85.59	84.52	93.28	82.70	82.90	82.41	91.79	82.38	82.07	80.84	89.94	69.28	70.14	68.67	78.77
Information-theoretic Feature	KNN	77.81	76.92	75.94	82.82	73.57	71.80	71.99	76.98	71.66	69.76	69.31	77.48	52.23	51.36	50.44	59.76
	SVM	84.85	84.20	83.35	88.70	74.39	75.39	73.50	84.53	74.57	72.85	72.51	80.67	51.62	51.59	46.76	60.46
	RF	84.57	82.91	82.85	89.73	74.30	74.74	73.77	80.36	76.75	76.53	74.37	80.38	59.89	59.34	57.60	65.17
	LDA	77.19	75.34	75.50	83.22	70.17	70.72	69.26	74.19	73.29	72.13	70.48	76.23	50.76	49.75	47.69	54.23
	XGBoost	86.72	86.07	85.12	92.62	78.02	77.55	77.28	85.41	78.38	77.37	76.59	88.59	58.57	57.96	54.47	62.78
	MLP	82.58	81.71	80.83	90.54	81.34	81.86	80.46	87.98	77.89	77.15	76.24	85.25	65.54	64.38	63.57	74.25
Frequency Feature	KNN	72.64	71.13	69.87	80.18	66.23	65.81	64.64	69.39	65.93	65.40	64.26	72.00	50.05	49.24	47.73	54.00
	SVM	79.40	78.70	77.81	87.88	68.67	68.65	67.51	75.09	68.20	67.88	65.35	79.97	52.57	50.27	49.12	57.70
	RF	80.48	80.18	78.94	87.82	73.78	72.31	72.29	79.55	72.97	71.69	70.15	79.78	58.53	57.92	54.46	68.02
	LDA	74.91	73.98	72.25	78.39	63.74	64.18	62.93	66.06	64.34	61.77	59.81	70.99	55.11	53.06	52.57	58.16
	XGBoost	82.12	79.42	79.62	91.74	73.02	70.45	71.70	79.53	78.96	78.34	77.30	85.82	59.82	59.26	57.65	68.50
	MLP	76.22	75.29	74.34	84.80	74.80	73.09	73.61	82.08	78.80	76.63	76.36	87.88	67.51	66.88	66.53	75.80
Temporal Fusion Features	KNN	83.88	82.62	82.13	91.83	76.16	76.41	75.31	81.10	76.83	76.29	74.62	84.14	52.33	52.20	50.28	57.31
	SVM	86.04	84.66	84.54	92.20	78.70	79.62	77.97	85.86	79.02	78.54	77.33	89.63	61.87	62.81	59.20	70.24
	RF	87.23	86.53	85.70	95.67	81.61	80.89	80.97	88.34	82.47	82.26	81.02	91.93	65.96	66.45	64.94	73.14
	LDA	81.78	80.53	80.07	90.50	71.11	71.61	69.97	75.87	75.75	74.73	73.44	84.86	59.45	59.93	56.32	69.57
	XGBoost	**89.02**	**88.36**	**87.61**	**96.06**	82.61	81.85	82.32	90.23	79.93	78.90	78.14	90.92	69.29	67.64	66.65	75.51
	MLP	88.48	87.01	87.11	95.75	**85.29**	**85.12**	**84.95**	**95.57**	**83.74**	**83.01**	**82.21**	**92.42**	**72.30**	**72.91**	**69.65**	**81.31**

Table 5
The classification results for MSPDnet with different number of BiMap/ReEig blocks on various evaluation criteria (Accuracy, Recall, F1-score, and AUC). The best results are labeled in bold.

		Accuracy	Recall	F1-score	AUC
UCLIC	MSPDnet-1	85.23	85.03	83.33	91.46
	MSPDnet-2	92.17	91.36	91.45	95.79
	MSPDnet-4	**93.69**	**92.53**	**92.12**	**96.89**
	MSPDnet-8	89.96	88.83	88.36	92.43
EGBM	MSPDnet-1	81.69	81.93	81.41	86.01
	MSPDnet-2	88.96	88.02	87.47	90.94
	MSPDnet-4	**89.74**	**88.81**	**88.33**	**94.37**
	MSPDnet-8	89.55	88.32	87.22	92.81
KDAE	MSPDnet-1	80.63	79.64	78.35	84.49
	MSPDnet-2	87.56	86.39	85.88	91.40
	MSPDnet-4	**90.76**	**89.72**	**89.51**	**95.65**
	MSPDnet-8	88.91	88.59	88.44	92.24
MPI	MSPDnet-1	75.73	74.67	72.22	80.63
	MSPDnet-2	81.81	80.83	79.70	85.47
	MSPDnet-4	84.34	83.24	82.56	**89.04**
	MSPDnet-8	**84.72**	**83.75**	**83.20**	86.97

emotional representations changes significantly, but it is still difficult to distinguish seven emotions. Finally, we also visualize the final mapped covariance matrix after the four BiMap/ReEig blocks and LogEig operations, as shown in Fig. 4(c) and (f), and it can be observed that the features of different emotions are linearly separable through multiple mapping layers of MSPDnet. The results further prove that the proposed MSPDnet can learn more discriminative emotion representations from posture covariance matrices.

Furthermore, we conducted experiments to evaluate the impact of different number of BiMap/ReEig blocks (1, 2, 4, and 8 blocks) on the performance of MSPDnet. The results are summarized in Table 5. The results show that increasing the number of BiMap/ReEig blocks can improve the MSPDnet performance within a certain range, but too many layers (8 layers) may lead to performance degradation. This decline in performance could be attributed to overfitting, especially when dealing with small-sample datasets like our affective body expression dataset. Therefore, we chose 4 layers as the optimal balance

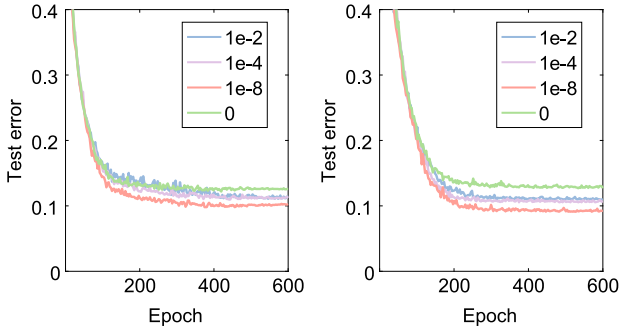


Fig. 5. The classification error curves of MSPDnet at different rectification thresholds ε on (a) EGBM dataset, and (b) KDAE dataset.

that guarantees the model's performance without increasing the model complexity excessively.

Fig. 5 compares the classification error for different rectification thresholds ε (1e−2, 1e−4, 1e−8, 0) on the EGBM and KDAE datasets, which can reflect the effect of different threshold choices on the convergence and performance of the network. Fig. 5 indicates that a higher ε may include too much noise in the features, while $\varepsilon = 0$ would eliminate the rectification effect, significantly reducing the model's performance. We found that ε = 1e−8 yielded the best convergence and performance across two datasets. Furthermore, the performance was relatively stable across different ε values, demonstrating the robustness of our model to this parameter.

4.5. Effectiveness of attentional temporal and spatial feature fusion

In Section 3.4, we introduce a novel fusion optimization algorithm, ATSFF, designed to integrate temporal and spatial features using a multiscale channel attention module. To confirm the effectiveness of ATSFF, we compared the classification performance of fused temporal–spatial features with temporal features (TF-BF) and spatial features before fusion (SF-BF). Furthermore, ATSFF was compared with advanced temporal and spatial fusion methods and the results are shown in Table 6.

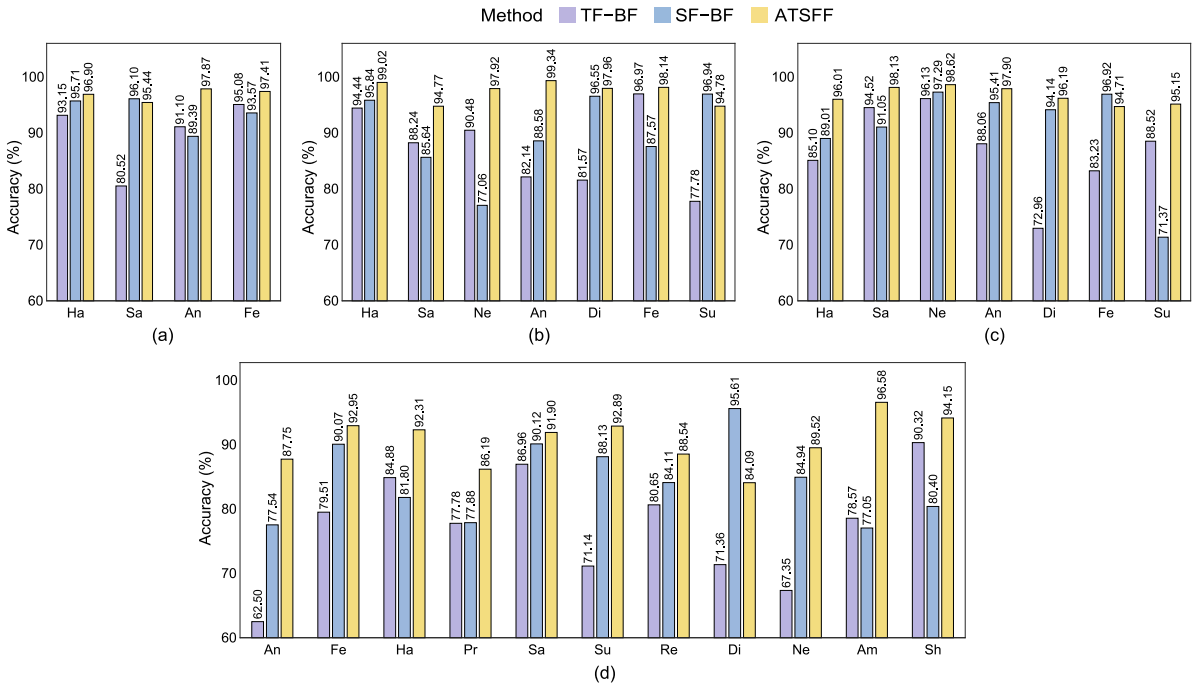


Fig. 6. The recognition accuracy (%) of temporal features, spatial features, and ATSFF for different emotions on (a) UCLIC dataset, (b) EGBM dataset, (c) KDAE dataset, and (d) MPI dataset.

Table 6
The classification results of different feature fusion methods on various evaluation criteria (Accuracy, Recall, F1-score, and AUC). The best results are labeled in bold.

	UCLIC				EGBM				KDAE				MPI			
	Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC	Accuracy	Recall	F1-score	AUC
TF-BF	89.98	89.00	88.50	94.52	87.33	86.38	85.75	94.09	86.92	85.97	85.32	94.45	77.07	76.20	75.09	83.35
SF-BF	93.69	92.53	92.12	96.89	89.74	88.81	88.33	94.37	90.76	89.72	89.51	95.65	84.34	83.24	82.56	89.04
TSFF-CON	92.38	91.29	91.39	94.63	91.90	90.93	90.02	93.30	91.31	90.28	90.07	94.29	81.35	80.17	79.45	85.26
TSFF-FC	94.89	93.85	93.73	96.16	93.87	92.61	92.32	94.96	93.37	92.41	92.09	95.59	85.93	84.95	84.29	91.73
TSFF-PWConv	94.87	93.78	93.71	97.14	92.84	91.70	91.98	94.34	92.53	91.55	91.18	96.19	86.97	86.01	85.40	92.55
ST-GCN [20]	91.84	90.85	90.43	94.95	93.19	92.19	91.84	97.89	91.90	90.99	90.66	96.88	83.85	82.92	82.17	91.19
MS-G3D [56]	95.20	94.18	93.92	98.25	94.36	95.33	94.51	96.00	93.48	92.60	92.32	97.96	89.37	88.42	87.99	94.41
ST-TR [57]	94.99	93.97	93.71	97.93	93.78	92.77	92.45	96.33	93.31	92.42	92.14	97.19	86.12	85.12	84.56	92.96
ATSFF	**96.91**	**95.80**	**95.67**	**98.90**	**97.43**	**97.25**	**96.99**	**98.44**	**96.67**	**95.94**	**94.96**	**98.27**	**90.61**	**89.50**	**88.98**	**96.49**

Table 7
The comparison of the proposed method and existing methods.

Research	Dataset	Acquisition device	Movement information	Methodology	Recognition performance
Burton et al. [58]	UCLIC	Mocap	Pos, Rot	LMA+HMM	72.00%
Wang et al. [59]	UCLIC	Mocap	Pos, Rot	Low/high-level postural and temporal features+RF	78.00%
Dewan et al. [60]	UCLIC	Mocap	Pos	LMA+LSTM	87.30%
Sapinski et al. [29]	EGBM	Kinect	Pos, Rot	RNN-LSTM	69.00%
Zhang et al. [25]	EGBM	Kinect	Pos, Rot	AS-LSTM	74.10%
Avola et al. [61]	KDAE	Mocap	Pos	Low-level postural features+MVRL	64.10%
Ghaleb et al. [28]	KDAE	Mocap	Pos, Rot	ST-GCNs	65.00%
Oguz et al. [32]	KDAE	Mocap	Pos, Rot	Postural statistical features+RegNetY-800MF	99.99%
Farinelli et al. [62]	KDAE	Mocap	Pos, Rot	LMA+LSTM	96.00%
Shirian et al. [63]	MPI	Mocap	Pos, Rot	GCN, L-GrIN	56.03% using GCN, 58.59% using L-GrIN
Crenn et al. [64]	UCLIC, MPI	Mocap	Pos, Rot	Posture and temporal features+SVM	UCLIC: 74.00%, MPI: 78.60%
Our method	UCLIC, EGBM, KDAE, MPI	Mocap, Kinect	Pos, Rot	Fused temporal–spatial feature	UCLIC: 96.91%, EGBM: 97.43%, KDAE: 96.67%, MPI: 90.61%

Abbreviations: AS-LSTM: LSTM network based on attention, GCN: Graph Convolution Network, HMM: Hidden Markov Model, L-GrIN: Learnable Graph Inception Network, LMA: Laban Movement Analysis, LSTM: Long Short-Term Memory, MVRL: Multi-View Representation Learning, Pos: Position, RegNetY-800MF: Regular Network Y-800 MegaFLOPs, RF: Random Forest, RNN: Recurrent Neural Network, Rot: Rotation, ST-GCNs: Spatio-Temporal Graph Convolutional Networks, SVM: Support Vector Machine.

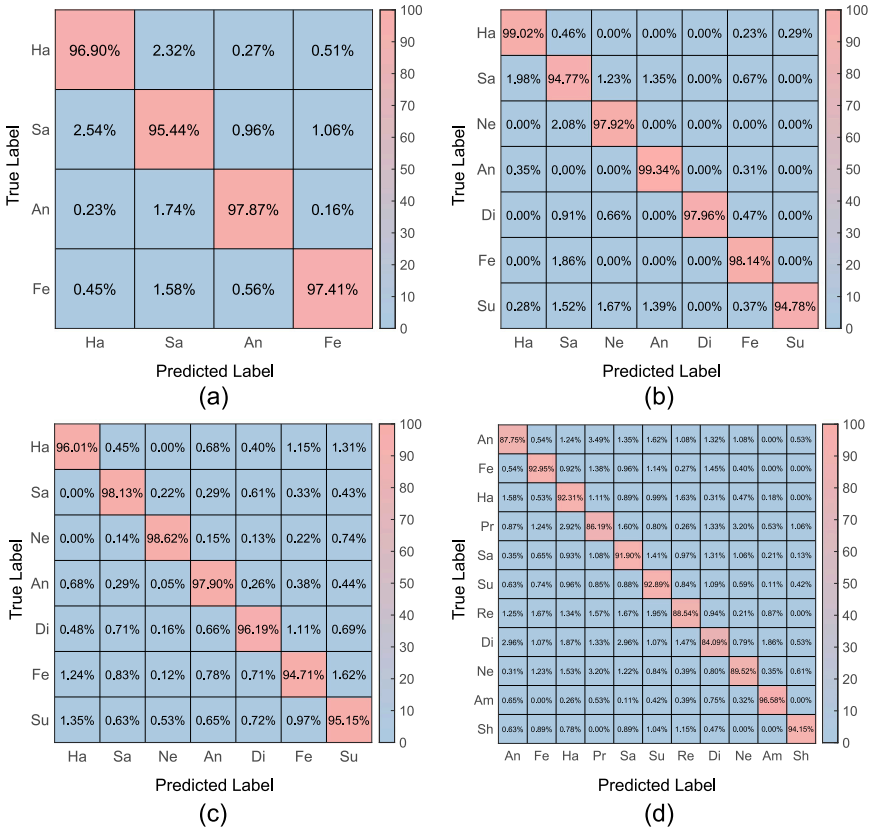


Fig. 7. The confusion matrices of the proposed method on the (a) UCLIC dataset, (b) EGBM dataset, (c) KDAE dataset, (d) MPI dataset.

As presented in Table 6, ATSFF outperformed individual temporal or spatial features across all datasets, with an accuracy improvement of 3.22–13.54%. Additionally, to assess the performance improvement of fused features on different emotions, Fig. 6 shows the recognition accuracies of individual temporal and spatial features compared to fused features for different emotions across four datasets. As shown in Fig. 6, the fused feature (ATSFF) well integrates the complementary between TF-BF and SF-BF in recognizing various emotions. For instance, in the KDAE dataset, TF-BF had only 88.52% accuracy and SF-BF showed 71.37% in recognizing surprise, while their fusion feature achieves 95.15% accuracy. These findings indicate that the proposed ATSFF significantly enhances the model's sensitivity and discrimination for different emotions.

To verify the superiority of the proposed two-branch ATSFF in jointly extracting global and local information, we conducted three ablation experiments to compare the proposed ATSFF with TSFF-CON, TSFF-FC and TSFF-PWConv. TSFF-CON is a traditional temporal–spatial feature concatenation method. TSFF-FC is a single-branch global feature fusion model constructed using GAP and FC, and TSFF-PWConv is a single-branch local feature fusion model constructed using PWConv. Furthermore, we also compare three advanced temporal and spatial fusion methods, i.e., spatial–temporal graph convolutional networks (ST-GCN) [20], multi-scale spatial–temporal graph convolution (MS-G3D) [56], and spatial–temporal transformer network (ST-TR) [57]. As shown in Table 6, ATSFF achieved the best recognition results on all datasets. These results indicate that ATSFF significantly outperforms traditional feature fusion methods and surpasses current advanced temporal–spatial fusion techniques. This may be attributed to ATSFF's two-branch network structure, which can dynamically extract contextual information of the fused features at both global and local scales, achieving attentional feature fusion across temporal and spatial domains.

4.6. Performance of the affective body expression recognition framework

Fig. 7 shows the recognition results of the proposed method on the four datasets in the form of confusion matrix. As shown in Fig. 7, the proposed method performs well in identifying all emotions on the UCLIC, EGBM, and KDAE datasets, achieving an accuracy rate exceeding 94% for each emotion. It is important to note that both UCLIC and EGBM are small sample datasets. The classification accuracy on the MPI dataset was lower compared to the other datasets. This may be due to the significant imbalance among the 11 emotional categories in the MPI dataset. Nevertheless, the model maintained a recognition accuracy exceeding 84% for all emotions, with particularly high accuracy rates for amusement and shame, surpassing 94%. These findings demonstrate the robustness of our method across datasets with significant variance, proving its effectiveness in handling small-scale and imbalanced data.

To evaluate the computational complexity of our proposed model, we calculated the Floating Point Operations (FLOPs) [65] of our model. FLOPs measure the number of floating-point arithmetic operations a model can perform in one second, providing a clear understanding of its efficiency and computational complexity. For the most complex KDAE dataset, the FLOPs value of our model is 0.22 G. This result demonstrates the lightweight nature of our model, which can achieve a good balance between accuracy and computational complexity.

4.7. Comparison with different methods

In this section, we present a comprehensive comparison of the proposed method with state-of-the-art methods. The comparison includes the used datasets, data acquisition equipment, anatomical movement information, specific methods and recognition results. As shown in Table 7, our method uses the most validation datasets, including major collection devices like Kinect and Mocap. Furthermore, the proposed

method effectively utilizes positional and rotational information from affective body expressions, achieving optimal classification results on the UCLIC, EGBM, and MPI datasets, with recognition accuracies of 96.91%, 97.43% and 90.61%, respectively. Specifically, our method outperforms the RF, SVM, and HMM based on hand-crafted posture features, as well as deep learning methods using LSTM and GCN architectures. These results further demonstrate the excellent recognition performance and generalization ability of our method.

On the KDAE dataset, our method also performed well, achieving a classification accuracy of 96.67%, although slightly lower than the results reported in [32]. It is important to note that the method in [32] achieved the best classification results on the KDAE dataset to date. In their work, the authors extracted several statistical features from body motions, including mean, root mean square (RMS), continuous wavelet transform (CWT), and joint neighborhood distance (JND). They then combined the extracted features with various machine learning and deep learning models for emotion recognition. However, this approach only considered statistical posture features in the temporal domain and achieved good results solely on the KDAE dataset. In contrast, our method effectively fused temporal–spatial features, achieving good classification performance across various datasets. This further demonstrates that the proposed method offers better generalizability in affective body expression recognition.

5. Conclusion and future work

This study introduces an affective body expression recognition framework to provide a universal and effective solution for decoding the complex mapping between emotions and body expressions. This framework extracts advanced and interpretable temporal–spatial features and employs a novel attentional feature fusion algorithm. Particularly, by introducing the interpretable body expression energy model (BEEM) and the multi-input symmetric positive definite matrix network (MSPDnet), our framework effectively quantifies and decodes the energy distribution, dynamic complexity, frequency activities, and the spatial Riemannian representation between body joints in body expressions. Furthermore, by utilizing a multiscale channel attention module with a two-branch bottleneck structure, the proposed ATSFF algorithm achieves attention temporal–spatial feature fusion at both global and local scales, significantly improving the performance of affective body expression recognition.

Although the proposed framework has made significant progress in emotional body expression recognition, there are still some limitations and constraints. First, a significant challenge for our approach is dealing with complex or imbalanced body expressions, such as those in the MPI dataset. To address this issue, future work will explore innovative data augmentation techniques. For instance, Generative Adversarial Networks (GANs) can be employed to generate more complex body expression data, thereby increasing the diversity of the training set. Additionally, geometric transformations and temporal perturbation techniques can be used to produce affective body expressions from different perspectives and scales. These techniques aim to create a more varied and representative training set, thus improving the model's generalizability and accuracy in decoding rare or complex body expressions. Second, the current recognition framework lacks research on the relative contribution of different body joints in body expressions. Future work will consider introducing modules of attention mechanisms targeting body joints within the recognition framework. These modules can utilize Self-Attention or Multi-Head Attention mechanisms to dynamically assign weights to body joints, enabling the model to focus more precisely on joint motions that carry key emotional information, thus improving emotion recognition performance.

Furthermore, we will continue to explore the potential applications of the proposed framework in various fields, including healthcare, virtual reality, and education. In healthcare, the framework could monitor patients' emotional states, providing valuable insights for mental health treatment and care. In virtual reality, enhancing avatars with the ability to recognize and express emotions through body expressions could lead to more immersive and engaging experiences for users. In education, recognizing students' emotions through body expressions could help educators tailor their teaching strategies to better meet the emotional needs of their students.

CRediT authorship contribution statement

Tao Wang: Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Shuang Liu:** Writing – original draft, Investigation, Conceptualization. **Feng He:** Validation. **Minghao Du:** Investigation. **Weina Dai:** Writing – review & editing. **Yufeng Ke:** Supervision. **Dong Ming:** Writing – review & editing, Supervision.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 81925020 and 81801786, and the General Program of Tianjin, China under Grant 19JCYBJC29200, and the Tianjin Research Innovation Project for Postgraduate Students under Grant 2022BKY053.

Data availability

Data will be made available on request.

References

[1] J.-Y. Huang, W.-P. Lee, Exploring the effect of emotions in human–machine dialog: an approach toward integration of emotional and rational information, Knowl.-Based Syst. 243 (2022) 108425.
[2] W. Jin, B. Zhao, Y. Zhang, J. Huang, H. Yu, WordTransABSA: enhancing aspect-based sentiment analysis with masked language modeling for affective token prediction, Expert Syst. Appl. 238 (2024) 122289.
[3] W. Jin, B. Zhao, L. Zhang, C. Liu, H. Yu, Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis, Inf. Process. Manage. 60 (3) (2023) 103260.
[4] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, R. Schettini, I. Shanin, Semi-supervised cross-lingual speech emotion recognition, Expert Syst. Appl. 237 (2024) 121368.
[5] K. Chen, X. Yang, C. Fan, W. Zhang, Y. Ding, Semantic-rich facial emotional expression recognition, IEEE Trans. Affect. Comput. 13 (4) (2022) 1906–1916.
[6] V. Padhmashree, A. Bhattacharyya, Human emotion recognition based on time–frequency analysis of multivariate EEG signal, Knowl.-Based Syst. 238 (2022) 107867.
[7] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey, IEEE Trans. Affect. Comput. 4 (1) (2013) 15–33.
[8] F. Noroozi, C.A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEE Trans. Affect. Comput. 12 (2) (2018) 505–523.
[9] A.P. Atkinson, W.H. Dittrich, A.J. Gemmell, A.W. Young, Emotion perception from dynamic and static body expressions in point-light and full-light displays, Perception 33 (6) (2004) 717–746.
[10] B. De Gelder, J. Van den Stock, The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole body expression of emotions, Front. Psychol. 2 (2011) 181.
[11] H.G. Wallbott, Bodily expression of emotion, Eur. J. Soc. Psychol. 28 (6) (1998) 879–896.
[12] H. Wang, A. Basu, G. Durandau, M. Sartori, A wearable real-time kinetic measurement sensor setup for human locomotion, Wearable Technol. 4 (2023) e11.
[13] H.J. Griffin, M.S. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, N. Bianchi-Berthouze, Laughter type recognition from whole body motion, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, 2013, pp. 349–355.

[14] S. Piana, A. Stagliano, F. Odone, A. Verri, A. Camurri, Real-time automatic emotion recognition from body gestures, 2014, arXiv preprint arXiv:1402.5047.
[15] D. Glowinski, M. Mortillaro, K. Scherer, N. Dael, G. Volpe, A. Camurri, Towards a minimal representation of affective gestures, in: 2015 International Conference on Affective Computing and Intelligent Interaction, ACII, IEEE, 2015, pp. 498–504.
[16] Y. Bhatia, A.H. Bari, G.-S.J. Hsu, M. Gavrilova, Motion capture sensor-based emotion recognition using a bi-modular sequential neural network, Sensors 22 (1) (2022) 403.
[17] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, R. Niewiadomski, Modeling multiple temporal scales of full-body movements for emotion classification, IEEE Trans. Affect. Comput. (2021) http://dx.doi.org/10.1109/TAFFC.2021.3095425, 1–1.
[18] S. Karumuri, R. Niewiadomski, G. Volpe, A. Camurri, From motions to emotions: classification of affect from dance movements using deep learning, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–6.
[19] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, D. Manocha, Step: Spatial temporal graph convolutional networks for emotion perception from gaits, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (02) 2020, pp. 1342–1350.
[20] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7444–7452.
[21] T. Wang, S. Liu, F. He, W. Dai, M. Du, Y. Ke, D. Ming, Emotion recognition from full-body motion using multiscale spatio-temporal network, IEEE Trans. Affect. Comput., 2023 (2023).
[22] Y. Zhai, G. Jia, Y.-K. Lai, J. Zhang, J. Yang, D. Tao, Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks, IEEE Trans. Affect. Comput., 2024 (2024).
[23] T. Wang, C. Li, C. Wu, C. Zhao, J. Sun, H. Peng, X. Hu, B. Hu, A gait assessment framework for depression detection using kinect sensors, IEEE Sens. J. 21 (3) (2020) 3260–3270.
[24] F. Ahmed, A.H. Bari, M.L. Gavrilova, Emotion recognition from body movement, IEEE Access 8 (2019) 11761–11781.
[25] H. Zhang, P. Yi, R. Liu, D. Zhou, Emotion recognition from body movements with as-lstm, in: 2021 IEEE 7th International Conference on Virtual Reality, ICVR, IEEE, 2021, pp. 26–32.
[26] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, A. Del Bimbo, Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices, in: International Conference on Image Analysis and Processing, Springer, 2017, pp. 550–560.
[27] U. Bhattacharya, C. Roncal, T. Mittal, R. Chandra, K. Kapsaskis, K. Gray, A. Bera, D. Manocha, Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping, in: European Conference on Computer Vision, Springer, 2020, pp. 145–163.
[28] E. Ghaleb, A. Mertens, S. Asteriadis, G. Weiss, Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–8.
[29] T. Sapiński, D. Kamińska, A. Pelikant, G. Anbarjafari, Emotion recognition from skeletal movements, Entropy 21 (7) (2019) 646.
[30] D. Avola, L. Cinque, A. Fagioli, G.L. Foresti, C. Massaroni, Deep temporal analysis for non-acted body affect recognition, IEEE Trans. Affect. Comput. 13 (3) (2020) 1366–1377.
[31] H. Zacharatos, C. Gatzoulis, P. Charalambous, Y. Chrysanthou, Emotion recognition from 3D motion capture data using deep CNNs, in: 2021 IEEE Conference on Games (CoG), IEEE, 2021, pp. 1–5.
[32] A. Oğuz, Ö.F. Ertuğrul, Emotion recognition by skeleton-based spatial and temporal analysis, Expert Syst. Appl. 238 (2024) 121981.
[33] B. Li, C. Zhu, S. Li, T. Zhu, Identifying emotions from non-contact gaits information based on microsoft kinects, IEEE Trans. Affect. Comput. 9 (4) (2018) 585–591.
[34] A. Kacem, M. Daoudi, B.B. Amor, S. Berretti, J.C. Alvarez-Paiva, A novel geometric framework on gram matrix trajectories for human behavior understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2020) 1–14.
[35] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2008, pp. 1–6.
[36] S. Saha, S. Datta, A. Konar, R. Janarthanan, A study on emotion recognition from body gestures using kinect sensor, in: 2014 International Conference on Communication and Signal Processing, IEEE, 2014, pp. 056–060.
[37] T. Wang, J. Sun, J. Chao, S. Zheng, C. Zhao, C. Wu, H. Peng, A novel gait analysis method based on the pseudo-velocity model for depression detection, in: 2020 IEEE International Conference on E-Health Networking, Application & Services, HEALTHCOM, IEEE, 2021, pp. 1–6.
[38] C. Ma, W. Li, J. Cao, J. Du, Q. Li, R. Gravina, Adaptive sliding window based activity recognition for assisted livings, Inf. Fusion 53 (2020) 55–65.
[39] P. Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoust. 15 (2) (1967) 70–73.
[40] G. Roffo, S. Melzi, M. Cristani, Infinite feature selection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4202–4210.
[41] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on riemannian manifolds, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1297–1304.
[42] Z. Huang, L. Van Gool, A riemannian network for spd matrix learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 2036–2042.
[43] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positive-definite matrices, SIAM J. Matrix Anal. Appl. 29 (1) (2007) 328–347.
[44] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, K. Barnard, Attentional feature fusion, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3560–3569.
[45] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[46] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, M.A. Chapman, Multi-scale point-wise convolutional neural networks for 3D object segmentation from LiDAR point clouds in large-scale environments, IEEE Trans. Intell. Transp. Syst. 22 (2) (2019) 821–836.
[47] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, pmlr, 2015, pp. 448–456.
[48] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 807–814.
[49] A. Kleinsmith, P.R. De Silva, N. Bianchi-Berthouze, Cross-cultural differences in recognizing affect from body posture, Interact. Comput. 18 (6) (2006) 1371–1389.
[50] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, G. Anbarjafari, Multimodal database of emotional speech, video and gestures, in: International Conference on Pattern Recognition, Springer, 2018, pp. 153–163.
[51] M. Zhang, L. Yu, K. Zhang, B. Du, B. Zhan, S. Chen, X. Jiang, S. Guo, J. Zhao, Y. Wang, et al., Kinematic dataset of actors expressing emotions, Sci. Data 7 (1) (2020) 1–8.
[52] E. Volkova, S. De La Rosa, H.H. Bülthoff, B. Mohler, The MPI emotional body expressions database for narrative scenarios, PLoS One 9 (12) (2014) e113647.
[53] J. Fang, T. Wang, C. Li, X. Hu, E. Ngai, B.-C. Seet, J. Cheng, Y. Guo, X. Jiang, Depression prevalence in postgraduate students and its association with gait abnormality, IEEE Access 7 (2019) 174425–174437.
[54] H.H. Lee, J.A. Emerson, D.M. Williams, The exercise–affect–adherence pathway: an evolutionary perspective, Front. Psychol. 7 (2016) 207868.
[55] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.
[56] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 143–152.
[57] C. Plizzari, M. Cannici, M. Matteucci, Skeleton-based action recognition via spatial and temporal transformer networks, Comput. Vis. Image Underst. 208 (2021) 103219.
[58] S.J. Burton, A.-A. Samadani, R. Gorbet, D. Kulić, Laban movement analysis and affective movement generation for robots and other near-living creatures, in: Dance Notations and Robot Motion, Springer, 2016, pp. 25–48.
[59] W. Wang, V. Enescu, H. Sahli, Adaptive real-time emotion recognition from body movements, ACM Trans. Interact. Intell. Syst. (TiiS) 5 (4) (2015) 1–21.
[60] S. Dewan, S. Agarwal, N. Singh, Laban movement analysis to classify emotions from motion, in: Tenth International Conference on Machine Vision (ICMV 2017), vol. 10696, SPIE, 2018, pp. 717–724.
[61] D. Avola, M. Cascio, L. Cinque, A. Fagioli, G.L. Foresti, Affective action and interaction recognition by multi-view representation learning from handcrafted low-level skeleton features, Int. J. Neural Syst. (2022) 2250040–2250040.
[62] L. Farinelli, Design and implementation of a multi-modal framework for scenic actions classification in autonomous actor-robot theatre improvisations, 2020.
[63] A. Shirian, S. Tripathi, T. Guha, Dynamic emotion modeling with learnable graphs and graph inception network, IEEE Trans. Multimed. 24 (2022) 780–790.
[64] A. Crenn, A. Meyer, H. Konik, R.A. Khan, S. Bouakaz, Generic body expression recognition based on synthesis of realistic neutral motion, IEEE Access 8 (2020) 207758–207767.
[65] J.L. Hennessy, D.A. Patterson, Computer Architecture: A Quantitative Approach, Elsevier, 2011.