Emotion Recognition From Full-Body Motion Using Multiscale Spatio-Temporal Network

Tao Wang, Shuang Liu, Feng He, Weina Dai, Minghao Du, Yufeng Ke, and Dong Ming, *Senior Member, IEEE,*

Abstract—Body motion is an important channel for human communication and plays a crucial role in automatic emotion recognition. This work proposes a multiscale spatio-temporal network, which captures the coarse-grained and fine-grained affective information conveyed by full-body motion and decodes the complex mapping between emotion and body movement. The proposed method consists of three main components. First, a scale selection algorithm based on the pseudo-energy model is presented, which guides our network to focus not only on long-term macroscopic body expressions, but also on short-term subtle posture changes. Second, we propose a hierarchical spatio-temporal network that can jointly process posture covariance matrices and 3D posture images with different time scales, and then hierarchically fuse them in a coarse-to-fine manner. Finally, a spatio-temporal iterative (ST-ITE) fusion algorithm is developed to jointly optimize the proposed network. The proposed approach is evaluated on five public datasets. The experimental results show that the introduction of the energy-based scale selection algorithm significantly enhances the learning capability of the network. The proposed ST-ITE fusion algorithm improves the generalization and convergence of our model. The average classification results of the proposed method exceed 86% on all datasets and outperform the state-of-the-art methods.

Index Terms—Emotion recognition, full-body motion, multiscale features, covariance matrix, Riemannian network, convolutional neural network.

1 Introduction

Automatic emotion recognition enables machines to communicate with humans in a natural and empathetic way, which is essential for improving efficiency and user experience in human-computer interaction (HCI) [1]. In recent decades, although many works have focused on vocal expressions, facial expressions and electroencephalography (EEG), emotion recognition based on full-body motion has not been extensively explored.

Full-body motion is the movement of the extremities, torso and other body parts, which is one of the most fundamental and natural non-verbal expression channels during affective communication [2]. Some special mapping relationships between body expressions and emotions have been reported; for example, the activities and dynamics of body movements are lower during low arousal emotions (e.g., sadness, relaxation) and higher during high arousal emotions (e.g., joy, anger) [3]. Furthermore, because of the large area of the human trunk, researchers can capture full-body motions in a nonintrusive manner at long distances, which makes it possible to recognize emotions in the wild [4]. Therefore, the modeling of full-body motion for emotion recognition has attracted the interest of researchers in the field of affective computing.

Full-body motions are performed in three-dimensional (3D) space; thus, 3D skeleton data are considered to be the most intuitive and effective method for representing body movements [5]. In addition, 3D skeleton data include rich spatial and temporal information about body movements [6], which allows us to explore the complex mapping relationship between emotions and full-body motions. With the development of inexpensive and portable depth sensors and real-time pose estimation algorithms [7], [8], we can easily and accurately acquire the 3D coordinates and rotation angles of each joint during full-body motion [9], which further promotes the research of emotion recognition based on 3D full-body skeleton data.

Previous research has shown that body expressions need to be analyzed over a certain length of time interval, and the full-body skeleton data on different time scales often contain different information. For example, the researchers of [10] found that long-term and macroscopic full-body motions need to be processed over long time intervals, whereas the subtle movements of some joints (e.g., hand tremors or trembling) can be identified over shorter time intervals. In [11], different time scales (ranging from 0.5 s to 5 s) were used to extract different levels of features from body skeleton sequences. The authors of [12] confirmed that multiscale analysis of full-body motions on different time scales is essential for emotion recognition. However, in previous studies, the different scales of body movement have been chosen randomly or empirically, and there is no gold standard for scale selection. This may lead to data on

- *Tao Wang, Shuang Liu, Minghao Du, and Yufeng Ke are with the Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China. E-mail: {taowang2021, shuangliu, duminghao98, clarenceke}@tju.edu.cn.*
- *Weina Dai is with the Department of Biomedical Engineering, College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin 300072, China. E-mail: woniudai9917@163.com.*
- *Feng He and Dong Ming are with the Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China, and also with the Department of Biomedical Engineering, College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin 300072, China. E-mail: {heaven, richardming}@tju.edu.cn.*

Manuscript received xx xx, 2022; revised xx xx, 2022.
Tao Wang and Shuang Liu contributed equally to this work.
(Corresponding authors: Shuang Liu and Dong Ming)

different time scales containing redundant or less affective information, thus reducing the identification performance. Furthermore, previous works mainly focused on analyzing spatial or temporal features independently [13], [14], and mostly used basic fusion techniques for spatio-temporal fusion, such as concatenation or summation operation [15]. However, these fusion methods mix information from different domains and limit the feature extraction capability.

To address the above problems, we propose a multiscale spatio-temporal network for automatic emotion recognition based on full-body motions, as shown in Fig. 1. Considering the lack of an effective and emotion-oriented scale selection method in existing multiscale modeling of affective body expression, this paper proposes an energy-based scale selection approach to guide our network to perform coarse-grained and fine-grained modeling simultaneously. Then, we construct a multiscale spatio-temporal network with a hierarchical two-branch architecture, which is a more efficient framework for spatio-temporal feature extraction and fusion. Specifically, to extract more discriminative spatio-temporal features, in the encoding of the spatio-temporal descriptors, inspired by the work of [13], [16], we construct the posture covariance matrix and the 3D posture image to encode full-body skeleton sequences in the spatial and temporal domains. Subsequently, we construct the decoding module of the multiscale spatio-temporal network based on the architecture of Riemannian network [17] and convolutional neural network (CNN) [18]. We redesign the structure of the two networks, including changing the layer arrangement and constructing the two-branch network structure, which can jointly process posture covariance matrices and 3D posture images with different time scales, and then hierarchically combine them in a coarse-to-fine manner. Finally, to jointly optimize the proposed spatial and temporal networks while alleviating overfitting during network fusion, we introduce an iteration-based fusion algorithm, namely spatio-temporal iterative (ST-ITE) fusion algorithm, which effectively reduces the complexity of the model while improving its generalizability and convergence.

The results of extensive experiments performed on five publicly datasets demonstrate the effectiveness and generalizability of the proposed method. Furthermore, our approach outperforms the existing state-of-the-art methods. The contributions of this paper can be summarized as follows:

- It designs an innovative energy-based scale selection algorithm to guide our network to learn not only macroscopic body expressions at long time scales, but also subtle posture changes at short time scales, thereby improving the learning ability of the network.
- It constructs a multiscale spatio-temporal network to decode the complex mapping between perceived emotions and full-body motions. Posture covariance matrices embedded with spatial correlation information and 3D posture images embedded with temporal dynamic information are used as inputs to the network.
- It proposes a spatio-temporal iterative (ST-ITE) fusion algorithm to enable the network to perform joint spatio-temporal optimization, reducing the complexity of the model while improving the generalizability and convergence of the model.

The rest of this paper is structured as follows. In Section 2, we provide a brief overview of related work. The datasets used in this paper are introduced in Section 3. In Section 4, we describe the proposed multiscale spatio-temporal network in detail. The experimental results of our approach are reported in Section 5. Finally, conclusion and future work is presented in Section 6.

2 Related Work

In the following section, we first review the literature on multiscale modeling of affective body expression. Then, we discuss research on the spatial and temporal analysis of body movement.

2.1 Multiscale Analysis of Affective Body Expression

Over the past decade, most research in affective body expression recognition have emphasized the importance of multiscale analysis of body movements. For instance, given that fine-grained features are critical for distinguishing actions, Kong et al. [19] developed multiscale temporal embedding modules to extract features at various temporal scales for skeleton-based action recognition. Recent work [20] also demonstrated the importance of different time scales in the analysis of full-body movements. The authors introduced a multiscale structure into traditional graph convolutional networks to extract multi-level posture information, which significantly improved the performance of the action recognition model. Wang et al. [21] proposed an action recognition framework based on the 3D CNN architecture that includes a module for modeling short-term to long-term temporal dependencies and can efficiently fuse multiscale features for action recognition. The work of [22] have demonstrated that the multiscale analysis can help models acquire more informative features in recognition tasks based on body movements.

However, in these studies, the multiscale data were selected randomly or empirically, which may generate redundant features and thus reduce recognition performance. Therefore, the question of how to establish an effective scale selection algorithm for multiscale modeling of affective full-body expression has attracted the interest of many researchers. The energy generated during postural movements is highly sensitive to different emotions [3], [23], which provides a new idea for multiscale analysis of affective full-body expressions. Some researchers believe that kinetic energy is a key indicator of full-body motions. For example, Glowinski et al. [24] regarded the total velocity of each joint in 3D space as kinetic energy and used a kinetic energy model to analyze affective body movements. Li et al. [25] suggested that a velocity model of 3D skeleton sequences calculated by coordinate differences can be used to explore the mapping relationships between emotions and postures. However, the energy generated by human postures can be expressed as both kinetic and potential energy. Gunes et al. [26] defined the first posture frame as a neutral frame and used the Euclidean distance between the

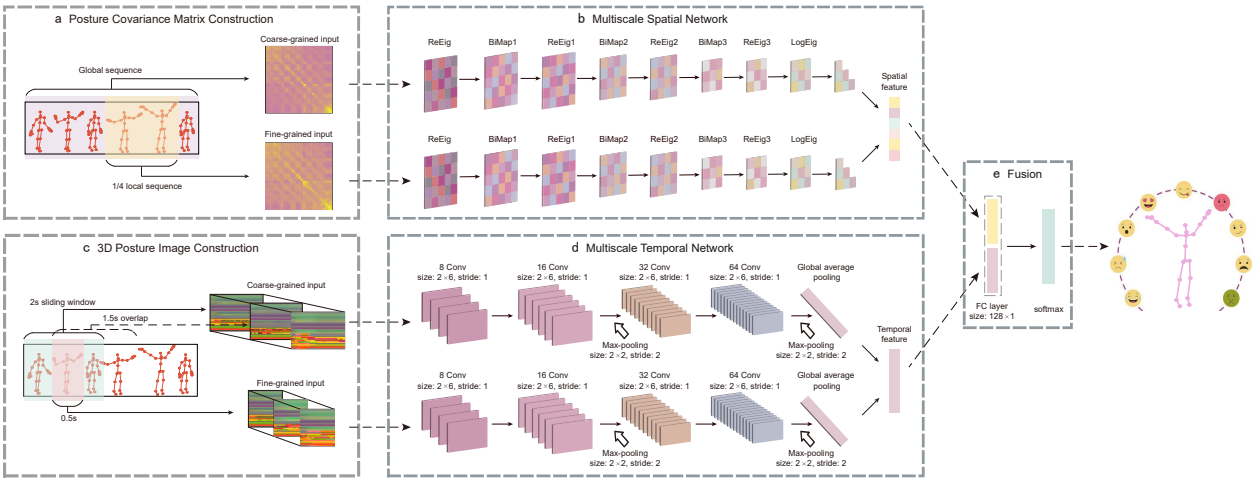

Fig. 1. The framework of the proposed method. For the spatial analysis, **(a)** the coarse-grained and fine-grained posture covariance matrices are constructed separately to encode global sequence and $1/4$ local sequence containing considerable emotional information; **(b)** a multiscale hierarchical spatial network based on the Riemannian network architecture is then proposed to jointly process input covariance matrices from coarse to fine. For the temporal analysis, **(c)** a 2s sliding window with 1.5s overlap is first utilized to divide the skeleton sequences into segments, in each segment, the whole sequence and the selected $1/4$ (0.5s) sequence are encoded in coarse-grained and fine-grained 3D posture images, respectively; **(d)** a multiscale hierarchical temporal network based on the CNN architecture is then proposed to jointly process input posture images from coarse to fine. Finally, **(e)** the spatial and temporal networks are jointly optimized in the FC layer by the proposed ST-ITE fusion algorithm and then fed into the softmax layer for emotion recognition.

neutral frame and the instantaneous posture as the potential energy feature to identify different emotions.

2.2 Spatial and Temporal Analysis of Affective Body Expression

For the spatial analysis of full-body skeleton sequences, existing methods often use posture covariance matrices to encode spatial correlations between skeleton joints, and exploit the geometric properties of the Riemannian manifold to extract features from the posture covariance matrices. For example, Daoudi et al. [13] represented 3D skeleton data with posture covariance matrices and exploited the Riemannian centre of mass and the log-Euclidean Riemannian metric to classify five emotions. Kacem et al. [27] proposed a geometric measure to process the posture covariance matrix for emotion recognition from full-body skeleton sequences. However, with significant advances in the study of optimization strategies and activation functions for Riemannian networks [28]–[30], an increasing number of researchers researchers have focused on the potential of Riemannian networks in processing posture covariance matrices. Huang et al. [17] used the Riemannian network to extract spatial information from the posture covariance matrix, significantly improving the performance of skeleton-based recognition tasks. Wang et al. [31] constructed a manifold-to-manifold Riemannian network, which can learn more discriminative low-dimensional features from the input posture covariance matrix.

For temporal analysis, most studies in the last decade have emphasized the effectiveness of encoding full-body skeleton sequences using image-based representations, while CNN have been proven to learn long-term temporal dependencies from the posture images [32], [33]. For instance, Ke et al. [34] transformed three channels of the skeleton sequence into the cylindrical coordinates to three clip images and utilized CNN to extract long-term temporal information from generated image representations. In their subsequent research [35], the authors further improved the original network into a hierarchical structure and proposed a multitask convolutional neural network (MTCNN) to process the generated clip images in parallel for skeleton-based action recognition. Laraba et al. [36] mapped the 3D coordinates of the joints in the pose skeleton sequence to red, green, and blue values in the RGB domain, resulting in an image-based representation, and then exploited discriminative features from the obtained images using CNN. Recently, image representations have been applied in emotion recognition based on full-body motion. For example, in the work of [16], the authors leveraged four graph coding formats to represent full-body skeleton data, and constructed multi-input CNN structures to process the four types of images. The experimental results indicate that the multi-input CNN shows great capability in learning emotion patterns from image representations.

3 Materials

The efficiency and generalizability of the proposed method was verified on five public affective full-body expression datasets, which are listed in Table 1. The five datasets were collected using different devices and included participants from diverse regions.

1) EGBM [37]: It was captured by the Kinect V2 sensor at a frame rate of 30Hz. This database contains 560 body motion samples of actors representing 7 emotions: happiness (Ha), sadness (Sa), neutral (Ne), anger (An), disgust (Di), fear (Fe), and surprise (Su). Each emotion is represented by 80 samples. The scenarios were performed by 16 professional Polish actors. Each segment contains the 3D position and orientation data of 25 joints. It should be noted that the actors' body movements during recording were not imposed or previously defined; thus, this database can be treated as a quasi-natural database.

2) KDAE [38]: It contains 3D body motions recorded at a frame rate of 125 Hz by a Noitom Perception Neuron (PN). This database contains 1402 full-body expressions representing 7 emotions: happiness (Ha), sadness (Sa), neutral (Ne), anger (An), disgust (Di), fear (Fe), and surprise (Su). The scenarios were performed by 22 Chinese actors, and each posture segment contains the position and rotation data of 72 anatomical nodes. In this paper, we use Right Hand and Left Hand (joints 19 and 47) instead of the hand position and remove other hand-related joints. After the above preprocessing steps, the total number of joints was reduced from 72 to 24.

3) Emilya [39]: It was recorded by the Xsens MVN system at a frame rate of 120 Hz. This database contains 8,206 emotional posture segments representing 8 emotions: anxiety (Ax), pride (Pr), happiness (Ha), sadness (Sa), panic fear (Fe), shame (Sh), anger (An), and neutral (Ne). The scenarios were performed by 12 actors representing 8 daily actions. Each posture segment contains the 3D position and rotation data of 28 markers.

4) MPI [40]: It was recorded by the Xsens MVN motion capture system at a sampling rate of 120 Hz. This database contains 1,447 body motion samples of actors representing 11 emotions: anger (An), fear (Fe), happiness (Ha), pride (Pr), sadness (Sa), surprise (Su), relief (Re), disgust (Di), neutral (Ne), amusement (Am), and shame (Sh). The scenarios were performed by 8 actors. Each posture segment contains the 3D position and rotation data of 28 markers. During the recording sessions, the actors were seated on a stool and asked to express different emotions through only upper body activities, so the 10 lower limb joints were excluded from the dataset. It should be noted that this database is highly imbalanced in terms of emotional expression, which is a challenge for the proposed model.

5) DMCD [41]: It was recorded by the PhaseSpace Impulse X2 MoCap system at a sampling rate of 120 Hz. This database contains 108 emotional dance sequences representing 12 emotions: afraid (Af), anger (An), annoyed (Ao), bored (Bo), excited (Ex), happiness(Ha), miserable (Mi), pleased (Pl), relaxed (Re), sadness (Sa), satisfied (St), and tired (Ti). The scenarios were performed by 6 dancers. Each posture segment contains the 3D position and rotation data of 38 markers. Similar to the preprocessing method in [12], we selected 26 joints for analysis. The subjects in this dataset performed complex dance movements to express different emotions, so it is a highly complex dataset of affective body expression.

TABLE 1
Description of the datasets used to evaluate the proposed method.

	EGBM [37]	KDAE [38]	Emilya [39]	MPI [40]	DMCD [41]
Acquisition Device	Kinect V2	Noitom PN	Xsens MVN	Xsens MVN	X2 MoCap
Frame Rate	30	125	120	120	120
Markers	25	72	28	28	38
Segments	560	1402	8206	1447	108
Emotions	7	7	8	11	12

4 METHODOLOGY

4.1 Multiscale Analysis

4.1.1 Pseudo-Energy Model

Previous studies have found that the energy generated from full-body movements is the most significant representation of the emotional state [42], [43]. Therefore, in this study, we assume that an energy model can quantify correlations between the body motions in each frame and emotions, thereby providing a basis for multiscale modeling of affective body expressions. The pseudo-energy model is defined by referring to the concept of mechanical energy.

The mechanical energy of the i^{th} joint in the f^{th} frame during full-body motions can be defined as follows:

$$E_i^f = E_{k,i}^f + E_{p,i}^f = \frac{1}{2} m \left(v_i^f\right)^2 + mgh_i^f \tag{1}$$

where $E_{k,i}^f$ and $E_{p,i}^f$ represent the kinetic energy and potential energy of the joint respectively. $i \in [1, I]$ and $f \in [1, F]$, where I is the total number of human joints and F is the total number of frames in posture segment. v_i^f is the velocity of the i-th joint in the f-th frame, which is obtained by computing the finite difference between the position of joint in frame f and $f-1$, and assume zero velocity at $f = 0$. h_i^f is the Euclidean distances between the i-th joint in the f-th frame and its corresponding neutral posture. The definition of a neutral posture is provided below. Fig. 2(a) and (b) visualize the kinetic energy information v and the potential energy information h in a posture representing raising hands with happiness.

It is assumed that all joints in the human body have the same mass m, and we ignore the gravitational acceleration g. Thus, the mechanical energy of the i-th joint in the f-th frame can be estimated as follows:

$$E_i^f = \frac{1}{2} \left(\frac{P_i^f - P_i^{f-1}}{\Delta f} \right)^2 + \left| P_i^f - P_{n,i} \right| \tag{2}$$

where $P_i^f = [x, y, z]$ represents the position of the i-th joint in the f-th frame, and Δf is one frame. The $P_{n,i}$ is the joint position of the neutral posture, which provides an initial position for full-body motions. By calculating the Euclidean distance between the motion position on each frame and the initial position (i.e., neutral pose), we can calculate the pseudo-potential energy generated from full-body motion in each frame. Similar to [44], in this study, the neutral posture is defined as a relaxed standing posture in which both arms rest on the side of the thighs and both legs are straight. In particular, we firstly define five vectors, namely, the vector corresponding to shoulder center and hip center (i.e. $\overrightarrow{l_1}$), the vectors corresponding to shoulders and wrists of both arms (i.e. $\overrightarrow{l_2}$ and $\overrightarrow{l_3}$), and the vectors corresponding to hip and ankle joints of both legs (i.e. $\overrightarrow{l_4}$ and $\overrightarrow{l_5}$). Then, we calculate the sum of the included angles between the above five vectors and the normal of the horizontal plane in each frame (i.e. $\alpha_{sum} = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5)$), and the 5-frame postures with the minimum α_{sum} are averaged to obtain a neutral posture, which is shown in Fig. 3(a). An illustration of instant postures and their neutral postures is given in Fig. 3(b).

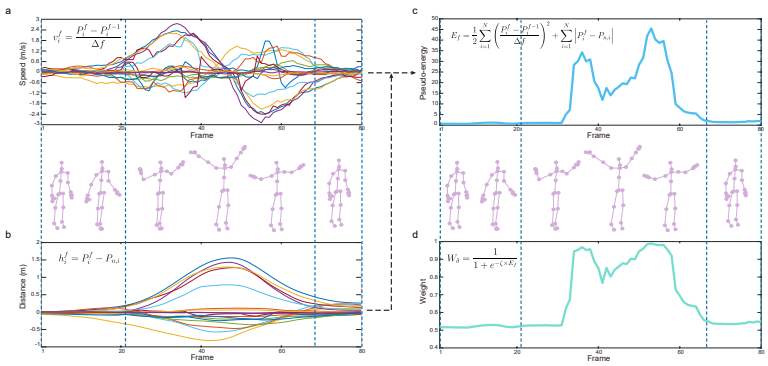

Fig. 2. The construction of the pseudo-energy model. **(a)** The kinetic energy information obtained by computing the finite difference between frames. **(b)** The potential energy information obtained by calculating the Euclidean distance between an arbitrary posture and its corresponding neutral posture. **(c)** The pseudo-energy model. **(d)** The emotional correlation weight.

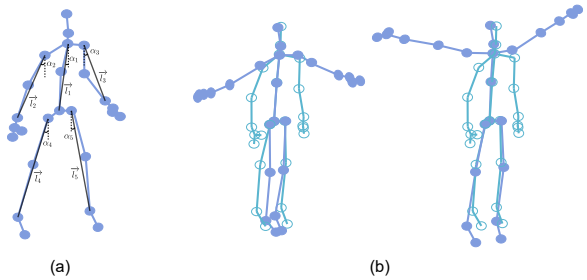

Fig. 3. Diagram of neutral posture. (a) Definition of neutral posture. (b) Diagrams of arbitrary postures (blue lines) and their corresponding neutral postures (cyan lines).

The obtained dynamic energy evolution of each joint over time is defined as the pseudo-energy model, which is illustrated in Fig. 2(c). Then we normalize the pseudo-energy model using Logistic function [44] to obtain the energy weight W_δ contained in the posture data of each frame, which is defined as follows:

$$W_\delta = \frac{1}{1 + e^{-\zeta \times E_f}} \tag{3}$$

$$E_f = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{P_i^f - P_i^{f-1}}{\Delta f}\right)^2 + \sum_{i=1}^{N}\left|P_i^f - P_{n,i}\right| \tag{4}$$

where E_f is the pseudo energy of all joints in the f^{th} frame. Given the energy generated during full-body movements is highly sensitive to the emotional state [3], [23], we hypothesize that the energy weight W_δ can represent the emotional information contained in the data of each frame, i.e., the W_δ is considered as the correlation weight between the skeleton data of each frame and emotion. Fig. 2(d) shows the correlation weights W_δ for a happy posture. As shown in Fig. 2(d), the full-body postures with larger weights have high motion activity and extension, such as swinging the body limbs and trunk. These frame data tend to contain more emotional information and distributed continuously at local locations in the sequence.

4.1.2 Scale Selection Algorithm

In this section, we design an adaptive scale selection algorithm based on the pseudo-energy model, which is shown in Algorithm 1. The algorithm can extract multiscale data containing sufficient emotional information by starting from the highest weight value in the pseudo-energy model and extending different lengths forward and backward. Specifically, we first utilize the pseudo-energy model to detect the frame with the highest emotional correlation weight in the posture sequence. Then, the scale window moves forwards and backwards, using the detected frame as the starting frame. If an inadequate extension length is detected on one side during the process, the algorithm automatically expands the window in the other direction until a sequence of preset lengths is collected, which ensures that the posture data in the window contains sufficient affective information.

In Algorithm 1, P_i represents the i-th full-body skeleton sequence, with starting and ending frames of S_i and E_i, respectively. n is the index value of different temporal scales, with $n \in [1, N]$. In addition, $\max(\cdot)$ returns the maximum element of an array, with a value of w_{max} and an index of f_{max}. $\text{round}(\cdot)$ represents rounding a value to its nearest integer. As shown in Algorithm 1, the scale selection algorithm outputs N full-body skeleton sequences with different temporal scales. In this paper, the number of scales N is set to 2. Subsequently, the scale selection algorithm is introduced into the proposed spatio-temporal network to guide our network to perform coarse-grained and fine-grained modeling simultaneously. In the multiscale spatio-temporal encoding, the scale size L_n for coarse-grained data is chosen as the length of the observation window, while the scale size for fine-grained data is chosen as $1/4$ of the length of the coarse-grained data. The specific encoding process for the multiscale spatial and temporal features can be found in Section 4.2.1 and 4.3.1.

4.2 Multiscale Spatial Feature Extraction

4.2.1 Posture covariance matrix Construction

Previous studies have shown that the posture covariance matrix can capture spatial geometric correlations between joints and has been applied as a robust and effective representation of full-body motion [13]. Therefore, in this section,

Algorithm 1 Scale Selection Algorithm

Input: sequence of full-body motions: $P_i = P\left[S_i,\ E_i\right]$
start frame of sequence: S_i
end frame of sequence: E_i
number of scales: N
size of scales: $L_1, .., L_n, .., L_N$
Output: start frame of scales: $S_1, .., S_n, .., S_N$
end frame of scales:$E_1, .., E_n, .., E_N$

initialization: $n = 1$
$W_\delta = \frac{1}{1+e^{-\zeta \times E_f(P_i)}}$
$[w_{max},\ f_{max}] = \max\left(W_\delta\right)$
while $n \leq \mathrm{N}$ **do**
  $S_n = f_{max} - \text{round}\left(\frac{1}{2} \times L_n\right)$
  $E_n = S_n + L_n - 1$
  if $S_n < S_i$ **then**
    $S_n = S_i$
    $E_n = S_n + L_n - 1$
  end if
  if $E_n > E_i$ **then**
    $S_n = S_n - E_n + E_i$
    $E_n = E_i$
  end if
  $n = n + 1$
end while

we introduce the posture covariance matrix to encode 3D full-body skeleton sequences.

Let $\mathbf{x} \in \mathbb{R}^d$ be d-dimensional feature vectors that contain the 3D position information of all body joints during body movements. A full-body skeleton sequence can be defined as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_f] \in \mathbb{R}^{d \times f}$, where $\mathbf{x}_f$ represents the position of the body joints in the f-th frame. Then, the posture covariance matrix of the posture sequence $\mathbf{X}$ is defined as:

$$\mathbf{C} = \frac{1}{f-1} \sum_{j=1}^{f} \left(\mathbf{x}_j - \boldsymbol{\mu}\right) \left(\mathbf{x}_j - \boldsymbol{\mu}\right)^{\mathrm{T}} \tag{5}$$

where $\boldsymbol{\mu}$ is the mean of $\mathbf{x}_f$.

Given the importance of multiscale spatial analysis in emotion recognition based on full-body motion, we extract $1/4$ of each global motion sequence as a local sequence. The local sequence is determined based on the proposed scale selection algorithm and it contains considerable emotional information. We then encode the global and local sequences of full-body motion with covariance descriptors. The two posture covariance matrices with different temporal scales are used as coarse-grained and fine-grained inputs in the subsequent multiscale spatial network. This procedure is illustrated in Fig. 1(a).

4.2.2 Multiscale Spatial Network

A non-singular posture covariance matrix belongs to the set of symmetric positive definite (SPD) matrices, which form a connected Riemannian manifold Sym_d^+ [45]. When the traditional neural networks based on Euclidean computations are used to process the posture covariance matrix, the non-euclidean input matrix must be vectorized during the mapping process of the network, resulting in the disappearance of spatial correlation between joints encoded in the matrix structure [46]. As an alternative, the Riemannian networks can capture more separable spatial features from the input SPD matrices by learning the manifold-to-manifold embedding mapping of the original matrix structure [17], [31]. Therefore, we propose a multiscale spatial network based on the Riemannian network architecture to learn the spatial affective representations encoded in the posture covariance matrices with different scales.

The proposed multiscale spatial network has two key characteristics. First, the network can directly process the posture covariance matrices without transforming the SPD matrices into vectors during the mapping process, which ensures that the spatial information encoded in the posture covariance matrices is not lost. Second, the network consists of two parallel shallow networks that jointly extract the spatial information embedded in the different scale posture covariance matrices. Specifically, one branch network learns the long-term macroscopic spatial patterns (coarse-grained modeling), while the other branch learns the short-term local spatial patterns (fine-grained modeling).

As shown in Fig. 4, the multiscale spatial network is composed of multiple parallel eigenvalue rectification (ReEig) layers, bilinear mapping (BiMap) layers, and eigenvalue logarithm (LogEig) layers. At the network input, $\mathbf{C}_0$ and $\mathbf{C}_0'$ represent coarse-grained and fine-grained posture covariance matrices, respectively. The ReEig layer rectifies the SPD matrix by using a non-linear function. To ensure that the input posture covariance matrix and the matrix mapped through the BiMap layer are still in Riemannian manifolds, we set a ReEig layer in the first layer of the network and after each BiMap layer. The BiMap layer transforms the input SPD matrix into a more discriminative matrix using the bilinear mapping transformation matrix. During this process, the input matrix does not need to be vectorized, thus preserving the spatial geometric information contained in the original SPD matrix. The LogEig layer endows elements in Riemannian manifolds with a Lie group structure, thereby reducing the matrix to a flat space in which traditional Euclidean computations can be applied [47].

Let $\mathbf{C}_{n-1} \in Sym_{d_{n-1}}^+$ be the input SPD matrix of size $d_{n-1} \times d_{n-1}$. The output $\mathbf{C}_{r,n}$ of the n-th ReEig layer, the output $\mathbf{C}_{b,n}$ of the n-th BiMap layer, and the output $\mathbf{C}_{l,n}$ of the LogEig layer can be defined as follows:

$$\mathbf{C}_{r,n} = f_r\left(\mathbf{C}_{n-1}, \varepsilon\right) = \mathbf{U}_{n-1}\ \mathrm{Max}\left(\varepsilon \mathbf{I}, \Lambda_{n-1}\right) \mathbf{U}_{n-1}^{\mathrm{T}} \tag{6}$$

$$\mathrm{Max}\left(\varepsilon \mathbf{I}, \Lambda_{n-1}\right) = \mathbf{E}(i,i) = \begin{cases} \Lambda(i,i), & \text{if } \Lambda(i,i) > \varepsilon \\ \varepsilon, & \text{if } \Lambda(i,i) \leq \varepsilon \end{cases} \tag{7}$$

$$\mathbf{C}_{b,n} = f_b\left(\mathbf{C}_{n-1}, \mathbf{W}_n\right) = \mathbf{W}_n \mathbf{C}_{n-1} \mathbf{W}_n^{\mathrm{T}} \tag{8}$$

$$\mathbf{C}_{l,n} = f_l\left(\mathbf{C}_{n-1}\right) = \mathbf{U}_{n-1} \log\left(\Lambda_{n-1}\right) \mathbf{U}_{n-1}^{\mathrm{T}} \tag{9}$$

where $\mathbf{U}_{n-1}$ and Λ_{n-1} denote the eigenvectors and eigenvalues of input matrix $\mathbf{C}_{n-1}$, respectively, and $\mathbf{I}$ is the

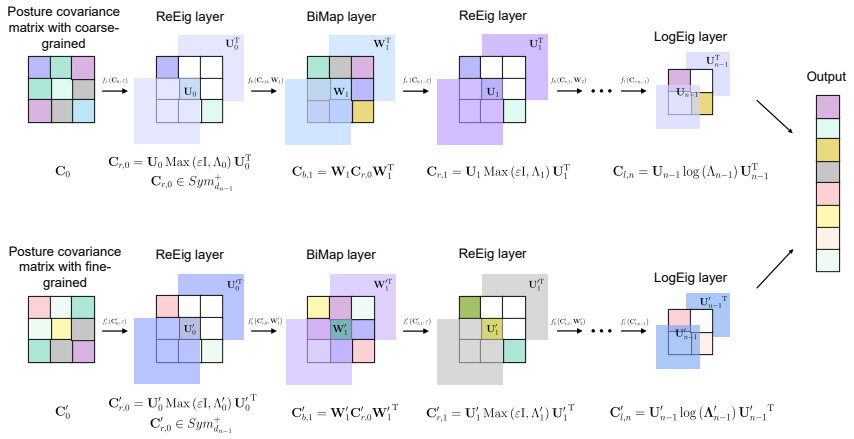

Fig. 4. Architecture of the proposed multiscale spatial network. $\mathbf{C}_0$ and $\mathbf{C}'_0$ represent coarse-grained and fine-grained posture covariance matrices, respectively. Each branch network consists of ReEig layers, BiMap layers and LogEig layers. The ReEig layer replaces the small eigenvalues with the preset threshold ε, thus ensuring the positive properties of mapped matrices. The BiMap layer transforms the input SPD matrix into a more discriminative matrix using the transformation matrix $\mathbf{W}$. The LogEig layer endows elements on Riemannian manifolds with a Lie group structure, so that the matrix can be reduced to a flat space. Finally, the outputs of the two branches are vectorized and concatenated into a 1-D vector. In each layer, we visually represent the matrix operations performed by the network using two light-colored matrices and one coloured matrix.

identity matrix. ε is a preset rectification threshold, which is used to replace null or small eigenvalues in Λ_{n-1} to obtain a new matrix $\mathbf{E}(i,i)$. $\mathbf{W}_n \in \mathbb{R}^{d_n \times d_{n-1}}$ is the bilinear mapping transformation matrix. The dimension of $\mathbf{W}_n$ is adjusted to reduce the complexity of the network, that is, $d_n < d_{n-1}$, and $\log(\cdot)$ is the matrix logarithm operation.

As shown in Fig. 1(b), because of the symmetry of the SPD matrix, for the output of each branch network, we extract only its lower triangle (including the diagonals) and reconstruct the result into a 1-D vector. Then, we concatenate the outputs of the two branches and regard it as the final output of the multiscale spatial network.

4.3 Multiscale Temporal Feature Extraction

4.3.1 3D posture image Construction

The above spatial features are computed over the whole posture segment, thus emphasizing the geometric correlations between joints during motion, while ignoring the temporal dynamic evolution of each joint. Therefore, inspired by the work of [16], [34], we propose a 3D posture image and combine it with the sliding window to encode the temporal dynamic evolution of affective full-body expressions in 3D space (X-, Y- and Z-axes). In detail, we first extract the 3D position and rotation information of each joint from the posture skeleton sequences. Then, we use the logistic position format mentioned in [16] to map the X, Y, and Z coordinate information to the R, G, and B components of an RGB image, respectively. The logistic position format maps the posture data to the interval -127 to +127 and is defined as follows:

$$R = \left\lceil \frac{255}{1 + e^{-L \times P}} \right\rceil \tag{10}$$

where R represents the value of the new joints after mapping, and L is an empirically chosen constant that is taken as 0.01 in this paper. P represents the relative position and rotation information obtained after body-centred normalization. Specifically, the spinebase (joint 0 in the EGBM dataset; the joint in the same position is used for the other datasets) in the first frame is treated as the origin of the local coordinate system, and the positions and rotations of all joints are taken with respect to this new origin. This operation is performed for each frame in each sequence of full-body motions. The motion trajectory of the i-th joint in the f-th frame after body-centred normalization is denoted as follows:

$$P_i^f = \begin{bmatrix} x_i^f - x_{\text{spinebase}}^1 \\ y_i^f - y_{\text{spinebase}}^1 \\ z_i^f - z_{\text{spinebase}}^1 \end{bmatrix}, \quad f \in \mathbb{N} \tag{11}$$

where x_i^f, y_i^f, and z_i^f represent the primitive 3D position and rotation parameters of the i-th joint in the f-th frame, and $x_{spinebase}^1$, $y_{spinebase}^1$, and $z_{spinebase}^1$ are the 3D parameters of the spinebase in the first frame. In this paper, we do not employ the traditional preprocessing method, i.e., using the 3D information of the spinebase on each frame for normalization, because this would cause the coordinate origin of the postural movement to be always located on the spinebase [9], resulting in the loss of displacement information of the trunk.

Finally, the position and rotation images are concatenated to obtain the 3D posture image. The encoding process is shown in Fig. 5. In the 3D posture image, all joints are represented on the vertical axis, while consecutive frames in the sequence of full-body motions are represented on the horizontal axis. Fig. 6 shows examples of 3D posture images for four emotions in the EGBM dataset.

As illustrated in Fig. 1(c), we apply a 2s sliding window with 1.5s overlap to divide the original full-body skeleton sequences into segments. In each segment of full-body motions, we utilize the proposed scale selection algorithm to select a quarter of local sequence that contain considerable emotional information. Then, the whole posture segment and the selected $1/4$ posture sequence are encoded into separate 3D posture images, which are then processed jointly in the

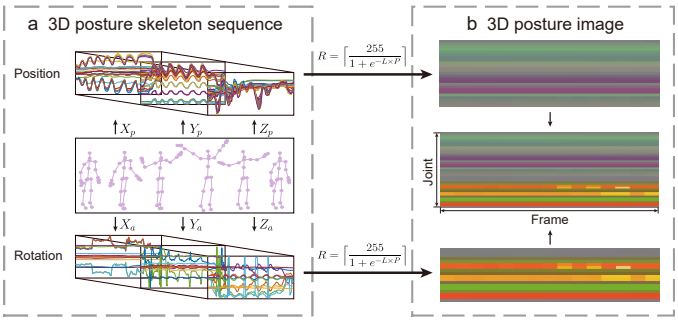

Fig. 5. The construction of the 3D posture image. **(a)** The process of extracting the 3D position and rotation information for each joint in the 3D full-body skeleton sequence. **(b)** The 3D posture image.

multiscale temporal network.

4.3.2 Multiscale Temporal Network

The CNN has been shown to be effective in developing temporal models of visual images [12]. Thus, the CNN is suitable for processing the proposed 3D posture image. Furthermore, compared with other neural networks, CNN has lower computational and memory costs and provides better performance on smaller datasets [16], which is suitable for emotional posture datasets with generally smaller datasets. Therefore, in this paper, we propose a multiscale temporal network based on the CNN architecture to jointly process 3D posture images with different scales.

In detail, similar to the above spatial network, the temporal network has a two-branch 2D-CNN architecture. One branch in the network performs coarse-grained modeling, while the other branch performs fine-grained modeling. The 3D posture images with different temporal scales are used as the inputs to the network. The proposed network is illustrated in Fig. 1(d). Each branch network consists of four convolution layers. The first convolution layer processes the input 3D gesture image with 8 filters. The following three layers have 16, 32 and 64 filters and all filter sizes are 2×6. Compared to common square filters (e.g., the 2×2 filter), the 2×6 rectangular filter allows our network to learn temporal features of the full-body expressions in consecutive frames (i.e., on the horizontal axis of the 3D posture image) rather than between skeleton joints (i.e., on the vertical axis of the 3D posture image). Finally, the outputs of the two branches are flattened and concatenated into a 1-D vector. This vector is fused with the obtained spatial feature in a fully connected (FC) layer for classification. For other settings of the multiscale temporal network, please refer to Section 5.1.

4.4 Spatio-Temporal Fusion Optimization Algorithm

Considering only spatial or temporal information is insufficient due to the intricate mapping between emotions and full-body motions. Therefore, in this section, the ST-ITE fusion algorithm is proposed to jointly optimize the above spatial and temporal networks. The features output by the spatial and temporal networks are fused by the FC layer and fed into a softmax layer to obtain the final prediction, as shown in Fig. 1(e). The optimization scheme is given in Algorithm 2. Given that spatial and temporal networks have a large number of parameters, and the emotional body expression datasets are typically small, this may easily lead to overfitting if both spatial and temporal networks are optimized at the same time. To address this problem, we apply an iterative algorithm to optimize the spatio-temporal network. Specifically, in each iteration, we select the network to be optimized by calculating the remainder of the iteration value t divided by 2. Then, we update the parameters of the selected network by using the gradient descent and backpropagation algorithms, during which the parameters of the other network are fixed. Finally, at the end of each iteration, the weights and biases of the FC layer are updated. With this fusion algorithm, the spatio-temporal network learns only half of the parameters in each optimization iteration, thereby reducing the complexity of the model while improving its generalisability and convergence.

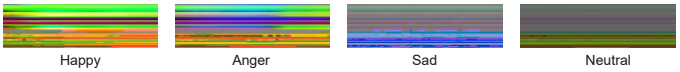

Fig. 6. The 3D posture images of four emotions on the EGBM dataset.

Algorithm 2 Spatio-Temporal Iterative (ST-ITE) Fusion Algorithm

Input: input of spatial network: X_{sc}, X_{sf}
input of temporal network: X_{tc}, X_{tf}
maximum number of iterations: i_{max}
learning rate: α_s, α_t, α_f
training set label: Y

Output: paremeters of spatial network: ω_{sc}, ω_{sf}
paremeters of temporal network: ω_{tc}, ω_{tf}
paremeters of FC layer: ω_f

initialization: iteration $i = i_s = i_t = 1$
while $i \leq i_{max}$ **do**
  if mod $(i, 2) = 1$ **then**
    $X_s^{i_s+1} = S(X_{sc},\ X_{sf},\ \omega_{sc}^{i_s},\ \omega_{sf}^{i_s})$
    $f^i = F\left(\left[X_s^{i_s+1},\ X_t^{i_t}\right], \omega_f^i\right)$
    $Loss = L\left(f^i, Y\right)$
    $\omega_{sc}^{i_s+1} = \omega_{sc}^{i_s} - \alpha_s(\frac{\partial Loss}{\partial f^i} \cdot \frac{\partial f^i}{\partial X_s^{i_s+1}} \cdot \frac{\partial X_s^{i_s+1}}{\partial \omega_{sc}^{i_s}})$
    $\omega_{sf}^{i_s+1} = \omega_{sf}^{i_s} - \alpha_s(\frac{\partial Loss}{\partial f^i} \cdot \frac{\partial f^i}{\partial X_s^{i_s+1}} \cdot \frac{\partial X_s^{i_s+1}}{\partial \omega_{sf}^{i_s}})$
    $i_s = i_s + 1$
  end if
  if mod $(t, 2) = 0$ **then**
    $X_t^{i_t+1} = T(X_{tc},\ X_{tf},\ \omega_{tc}^{i_t},\ \omega_{tf}^{i_t})$
    $f^i = F\left(\left[X_s^{i_s},\ X_t^{i_t+1}\right], \omega_f^i\right)$
    $Loss = L\left(f^i, Y\right)$
    $\omega_{tc}^{i_t+1} = \omega_{tc}^{i_t} - \alpha_t(\frac{\partial Loss}{\partial f^i} \cdot \frac{\partial f^i}{\partial X_t^{i_t+1}} \cdot \frac{\partial X_t^{i_t+1}}{\partial \omega_{tc}^{i_t}})$
    $\omega_{tf}^{i_t+1} = \omega_{tf}^{i_t} - \alpha_t(\frac{\partial Loss}{\partial f^i} \cdot \frac{\partial f^i}{\partial X_t^{i_t+1}} \cdot \frac{\partial X_t^{i_t+1}}{\partial \omega_{tf}^{i_t}})$
    $i_t = i_t + 1$
  end if
  $\nabla\omega_f^{i+1} = \frac{\partial Loss}{\partial f^i} \cdot \frac{\partial f^i}{\partial \omega_f^i}$
  $\omega_f^{i+1} = \omega_f^i - \alpha_f \nabla\omega_f^{i+1}$
  $i = i + 1$
end while

In Algorithm 2, X_{sc} and X_{sf} represent the full-body skeleton data fed into the coarse-grained and fine-grained branches in the spatial network, respectively. ω_{sc} and ω_{sf} are the weight parameters of the two branch networks. The parameter definitions of the temporal network are similar. In addition, α_s, α_t and α_f represent the learning rates of the spatial network, temporal network and FC layer, respectively. i_s denotes the i_s-th parameter update in the spatial network; similarly, i_t is the i_t-th optimization of the temporal network. i represents the i-th iteration of the entire spatio-temporal network. In the optimization process of the spatial network, mod $(i, 2)$ returns the remainder after dividing the iteration number i by 2. The mapping operation of the Riemannian network is represented by $S(\cdot)$. The fused spatio-temporal feature is denoted by $\left[X_s^{i_s+1},\ X_t^{i_t}\right]$, where $X_s^{i_s+1}$ is the spatial feature captured in the i_s-th optimization of the spatial network, which is fused with the previously obtained temporal feature $X_t^{i_t}$ captured in the i_t-th optimization of the temporal network. $F\Big(\left[X_s^{i_s+1},\ X_t^{i_t}\right], \omega_f^i\Big)$ denotes the fusion mapping of the spatio-temporal features in the FC layer, and ω_f^i denotes the weight parameters of the i-th iteration in the FC layer. The temporal network is optimized in a similar manner. The cross-entropy loss function $L\left(f^i, Y\right)$ is used during the network optimization, which has been widely used for various classification tasks and is known for its good convergence properties [48], [49]. This ensures that our ST-ITE fusion algorithm exhibits excellent convergence and stability.

5 Results and Discussion

In this section, we evaluate the efficiency of the proposed method on five public databases, which are described in Section 3. First, we discuss the effect of the multiscale analysis on emotion recognition based on full-body motion (Section 5.2). Then, the contribution of the proposed ST-ITE fusion algorithm is investigated (Section 5.3). Next, we report the performance of the proposed emotion recognition model (Section 5.4). Finally, we compare the results of our method with those of state-of-the-art approaches (Section 5.5).

5.1 Implementation Details

In this paper, for the spatial network, each branch network contains 3BiRe, which means that 3 blocks of BiMap/ReEig are used. For example, the structure of a branch network is $\boldsymbol{X}_0 \rightarrow f_r \rightarrow f_b^{(1)} \rightarrow f_r^{(1)} \rightarrow f_b^{(2)} \rightarrow f_r^{(2)} \rightarrow f_b^{(3)} \rightarrow f_r^{(3)} \rightarrow f_l$, where f_r, f_b, f_l denote the ReEig, BiMap, and LogEig respectively. The rectification threshold ε in the ReEig layer is set to 1e-8. The sizes of three transformation matrices $\mathbf{W}$ in the BiMap layers are set to $d_{n-1} \times 50$, 50×30, and 30×10 respectively, where d_{n-1} is the dimension of the input posture covariance matrix. For the temporal network, we design a CNN with four convolutional layers. The number of filters in each layer is 8, 16, 32, and 64, respectively. All filter sizes are 2×6, and all strides during the convolution are set to 1 with the "same" padding. The rectified linear unit (ReLU) activation function is used. After the convolution filters in layers 2 and 4, we introduce max-pooling operations with a kernel size of 2×2 and a stride of 2. The features output by the spatial and temporal networks are fused by a FC layer and fed into a softmax layer to obtain the final prediction. The FC layer with a layer size of 128 hidden nodes is used, and the loss was estimated with the cross-entropy loss function. The optimizer is Adam, with a learning rate of 10^{-3}. The batch size is set to 64, and the number of epochs is set to 200. In addition, the early stopping trick is adopted to prevent overfitting. The proposed network is trained using Tensorflow on two NVIDIA 3090 GPUs. All the experiments are performed using a 10-fold cross-validation scheme, and the classification performance is evaluated using the average classification accuracy[1].

5.2 Benefit of Multiscale Features

To validate the benefits of multiscale analysis in the proposed model, we first compared the classification performance when using single scale features (coarse-grained or fine-grained features) and multiscale features. Table 2 presents the average classification accuracy of different scale features on the five datasets. In addition, the paired sample t-tests were used to evaluate significant differences between the best result and other results. As shown in Table 2, the multiscale features (i.e., Multiscale (SS-PEM)) achieved higher accuracy, with a performance improvement of 1.69-12.90% over the single-scale features (i.e., Coarse-grained and Fine-grained features) on all datasets (Emilya: $p < 0.01$, other datasets: $p < 0.001$). This result demonstrates that the multiscale analysis is beneficial for emotion recognition based on full-body motion.

As presented in Section 4.1, the scale selection algorithm based on the pseudo-energy model was designed to enhance the learning ability of our network in a coarse-to-fine manner. To verify the effectiveness of the proposed scale selection algorithm, we compared the recognition performance of different scale selection methods, and the results are presented in Table 2. SS-BR in Table 2 represents the empirical scale selection method proposed in the work of [12], which uses the entire data and the last quarter of data in each skeleton segment as the multiscale features. Moreover, SS-KE and SS-PE represent scale selection methods based on kinetic energy and potential energy, respectively. The proposed scale selection algorithm based on the pseudo-energy model is abbreviated as SS-PEM. As shown in Table 2, on all datasets, the SS-PEM achieved better results than the other scale selection methods. The results imply that the energy-based scale selection scheme is more suitable for multiscale analysis of body expressions than other methods.

In Section 4.2.1 and 4.3.1, we chose the length of fine-grained data to be 1/4 of the length of coarse-grained data, i.e., 25%. To investigate the impact of the ratio between coarse-grained and fine-grained data lengths on the results, we computed the results for three different ratios: 10%, 50%, and 75%. The results are shown in the last four rows of Table 2. With the exception of the EGBM dataset, the results for the other four datasets show that the optimal classification results for multiscale features are achieved when the length of fine-grained data is 25% of the coarse-grained data. This finding reveals that shorter fine-grained data may make

1. Interested readers can contact the authors for access to the code, and we will be happy to provide the necessary resources.

TABLE 2
The average accuracies (%) of different scale analysis methods on the five datasets. The best results are labeled in bold.

	EGBM	KDAE	Emilya	MPI	DMCD
Coarse-grained	$86.88 \pm 2.71^{**}$	$87.54 \pm 1.60^{**}$	$92.73 \pm 1.49^{*}$	$80.05 \pm 2.09^{**}$	$78.87 \pm 3.97^{**}$
Fine-grained	$84.24 \pm 2.21^{**}$	$84.10 \pm 1.10^{**}$	$90.59 \pm 0.85^{**}$	$76.70 \pm 2.26^{**}$	$80.30 \pm 4.99^{**}$
SS-BR	$89.10 \pm 2.52^{**}$	$81.64 \pm 1.62^{**}$	$91.16 \pm 1.13^{**}$	$81.61 \pm 1.55^{**}$	$82.94 \pm 3.10^{*}$
SS-KE	$91.96 \pm 1.62^{**}$	$92.29 \pm 0.99^{**}$	93.65 ± 0.73	$85.26 \pm 1.22^{**}$	$83.31 \pm 4.17^{*}$
SS-PE	$91.64 \pm 2.72^{**}$	$91.94 \pm 0.88^{**}$	$92.18 \pm 1.07^{**}$	$83.85 \pm 1.74^{**}$	$82.30 \pm 3.12^{*}$
Multiscale (75%)	$92.14 \pm 1.65^{**}$	$91.32 \pm 0.75^{**}$	$90.75 \pm 0.68^{**}$	$85.04 \pm 2.04^{**}$	$72.88 \pm 5.69^{**}$
Multiscale (50%)	$\mathbf{96.19 \pm 1.93}$	$91.55 \pm 1.41^{**}$	$91.15 \pm 0.91^{**}$	$83.48 \pm 1.83^{**}$	$81.94 \pm 2.90^{**}$
Multiscale (10%)	$86.10 \pm 3.21^{**}$	$92.12 \pm 1.57^{**}$	$91.67 \pm 1.87^{**}$	$85.32 \pm 1.47^{**}$	85.49 ± 6.07
Multiscale (SS-PEM, 25%)	95.55 ± 1.47	$\mathbf{95.60 \pm 0.62}$	$\mathbf{94.42 \pm 0.68}$	$\mathbf{89.60 \pm 1.64}$	$\mathbf{86.15 \pm 3.02}$

* There are significant differences between the best results and other results ($^{*}: p < 0.01$, $^{**}: p < 0.001$).

multiscale features contain less information, thus reducing the classification performance of the model. On the other hand, when the length of the fine-grained data exceeds 25% of the coarse-grained data, the multiscale features may contain redundant information, thereby reducing the descriptiveness of multiscale features.

5.3 Effectiveness of Spatio-Temporal Fusion Optimization Algorithm

In Section 4.4, we propose the ST-ITE fusion optimization algorithm to jointly optimize the spatio-temporal network. To confirm the effectiveness of our algorithm, we first compared the classification performance of the fused features with the performance when only spatial or temporal features were used. Then, we compared the proposed ST-ITE fusion algorithm with the conventional optimization algorithm (ST-SIM) that simultaneously optimizes the spatial and temporal networks in each iteration. These results are shown in Table 3. On all five datasets, the accuracy obtained by the fused features (e.g., ST-SIM and ST-ITE) was significantly higher than the accuracies obtained when using either spatial or temporal features, with a performance improvement of 4.37-18.54% (all: $p < 0.01$). This result shows that the fused spatio-temporal feature is more discriminative for emotion recognition. Furthermore, compared with ST-SIM, the proposed ST-ITE achieved better performance with smaller standard deviation (Emilya: $p < 0.05$, other datasets: $p < 0.01$), demonstrating that the ST-ITE fusion algorithm effectively improved the performance of the model and alleviated the impact of individual differences.

Fig. 7 illustrates the recognition accuracy of different emotions on the five datasets using the features before fusion, and the fused features obtained using ST-SIM and ST-ITE. Compared with the spatial features, temporal features, and the features obtained by ST-SIM, the spatio-temporal features obtained by ST-ITE achieved better recognition results for the majority of emotions on the five datasets. For instance, the recognition accuracies of neutral and anger on the EGBM dataset were increased by 8.28% and 7.05%, respectively. The recognition accuracies of relief and happiness on the MPI dataset were improved by 9.74% and 7.81%, respectively. This finding demonstrates that the proposed ST-ITE fusion algorithm significantly improves the performance of the model in identifying different emotions. In addition, the performance of the fused features obtained by ST-ITE is relatively balanced for different emotions.

Fig. 8 presents the accuracy curves on the testing set before and after feature fusion on the EGBM and KDAE datasets. On the EGBM and KDAE datasets, the accuracies of ST-ITE exceeded 94% after 65 and 50 epochs, and its convergence performance was significantly better than ST-SIM and the features before fusion. These results suggest that our ST-ITE fusion algorithm improves the convergence of the spatio-temporal network.

In addition, we further compared the proposed fusion algorithm with three advanced spatio-temporal fusion algorithms, and the results are shown in Table 3. Since few existing works introduce spatio-temporal joint learning in affective body expression recognition, we compare three action recognition methods with advanced spatio-temporal fusion algorithms [50]–[52], which are similar to the emotion recognition task based on body movements. We obtained the original code for these three methods and evaluated their performance on all five affective body expression datasets. As shown in Table 3, the proposed ST-ITE fusion algorithm achieved optimal recognition results on four datasets, with performance improvements of 0.23-5.75%. On the DMCD dataset, our method only has a lower performance than MS-G3D, which represents to date a very excellent spatio-temporal fusion method in skeleton-based recognition tasks. The results imply that the proposed spatio-temporal fusion algorithm is more suitable for emotion recognition based on full-body motion than other traditional spatio-temporal learning methods.

5.4 Performance of the Emotion Recognition Model

Fig. 9 shows the results of the proposed method in the form of a confusion matrix for the five datasets, where each row in the confusion matrix represents a ground truth class and each column represents a predicted class. As presented in Fig. 9, on the EGBM, KDAE, and Emilya datasets, the proposed method performed well in recognizing all emotions, and the accuracies for each emotion exceeded 90%. It is important to note that these three datasets were collected by different devices, and the participants were from various countries. In addition, the EGBM dataset is small, and the affective body expressions included in this dataset were

TABLE 3
The average accuracies (%) before and after feature fusion on the five datasets and the comparison with the existing spatio-temporal fusion methods. The best results are labeled in bold.

	EGBM	KDAE	Emilya	MPI	DMCD
Spatial feature	$88.24 \pm 2.04^{**}$	$87.97 \pm 0.90^{**}$	$75.88 \pm 0.69^{**}$	$72.39 \pm 1.38^{**}$	$79.75 \pm 2.54^{**}$
Temporal feature	$81.88 \pm 3.78^{**}$	$87.14 \pm 1.27^{**}$	$90.05 \pm 1.40^{**}$	$83.51 \pm 1.26^{**}$	$77.13 \pm 1.91^{**}$
ST-SIM	$91.78 \pm 1.94^{**}$	$92.89 \pm 1.56^{**}$	$93.65 \pm 0.73^{*}$	$84.00 \pm 1.60^{**}$	$83.27 \pm 3.85^{**}$
ST-GCN [50]	$93.19 \pm 1.73^{*}$	$91.90 \pm 1.51^{**}$	$92.73 \pm 1.49^{**}$	$83.85 \pm 1.74^{**}$	$84.25 \pm 3.48^{**}$
MS-G3D [51]	$94.36 \pm 1.31^{*}$	$93.48 \pm 0.72^{**}$	93.82 ± 0.70	89.37 ± 0.97	$\mathbf{88.10 \pm 1.96}$
ST-TR [52]	$93.78 \pm 2.06^{*}$	$93.31 \pm 0.64^{**}$	$91.24 \pm 0.86^{**}$	$86.12 \pm 1.42^{**}$	$84.71 \pm 2.44^{*}$
ST-ITE	$\mathbf{95.55 \pm 1.47}$	$\mathbf{95.60 \pm 0.62}$	$\mathbf{94.42 \pm 0.68}$	$\mathbf{89.60 \pm 1.64}$	86.15 ± 3.02

* There are significant differences between the best results and other results ($^{*}: p < 0.05$, $^{**}: p < 0.01$).

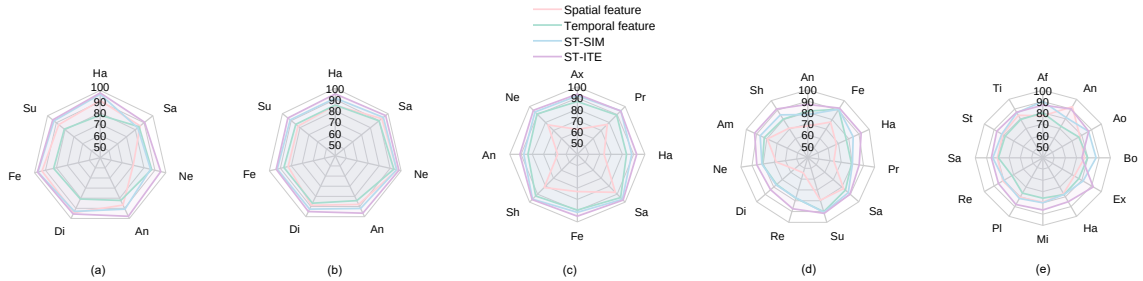

Fig. 7. The recognition accuracies (%) of the spatial feature, temporal feature, ST-SIM, and ST-ITE for different emotions on the (a) EGBM dataset, (b) KDAE dataset, (c) Emilya dataset, (d) MPI dataset, and (e) DMCD dataset.

freely performed by actors without any additional restrictions, which could result in different representations during each repetition. These factors were all challenges for our model. However, the proposed method still achieved excellent results on the above three datasets, which demonstrates the robustness and generalizability of our approach.

As presented in Fig. 9(d) and (e), the proposed method had lower classification accuracies on the MPI and DMCD datasets than on the other datasets. This result could be because the MPI dataset is highly imbalanced among the 11 emotional classes, which causes the trained model to be biased towards the majority class [53]. Furthermore, the DMCD dataset is a highly complex dataset, which may increase the difficulty of the model in extracting multiscale spatio-temporal features. Nevertheless, our model achieved a remarkable recognition rate of more than 83% for all emotions on the MPI dataset. On the DMCD dataset, the proposed method achieved a recognition performance of over 85% for 11 emotions except "bored". These results suggest that the proposed multiscale spatio-temporal network is robust to data with class imbalance and high complexity.

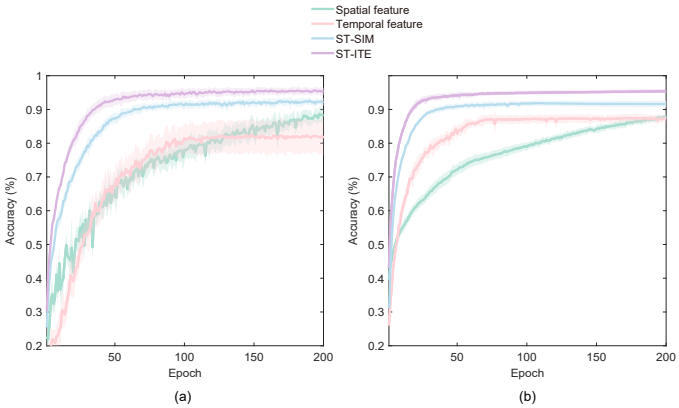

Fig. 8. The accuracy curves on the testing set before and after feature fusion on the (a) EGBM dataset and (b) KDAE dataset.

5.5 Comparison with State-of-the-Art Methods

In this section, we compared our results with state-of-the-art methods on the five datasets. To allow a fair comparison, we adopted the cross validation settings mentioned in the corresponding literature for the different datasets. For example, for the EGBM dataset, we employed a 10-fold cross-validation and a leave-one-subject-out (LOSO) protocol. The methods used in the comparison are described in the following:

1) For the EGBM dataset, Sapinski et al. [54] extracted a sequence of key frames from the full-body motions and used a CNN, a recurrent neural network (RNN) and an RNN with long short-term memory network (RNN-LSTM) to perform affective body expression recognition. Zhang et al. [55] proposed an attention-based stacked LSTM network (AS-LSTM) for emotion recognition from body movements.

2) For the KDAE dataset, Avola et al. [56] proposed a pipeline using multi-view representation learning (MVRL) for affective action recognition. Ghaleb et al. [57] represented the posture sequence as a graph, which was fed into the spatio-temporal graph convolutional networks (ST-GCNs) for emotion recognition.

3) For the Emilya dataset, in [59], 114 hand-crafted posture features were extracted, and the random forest (RF) with 500 trees was used to process the obtained features for

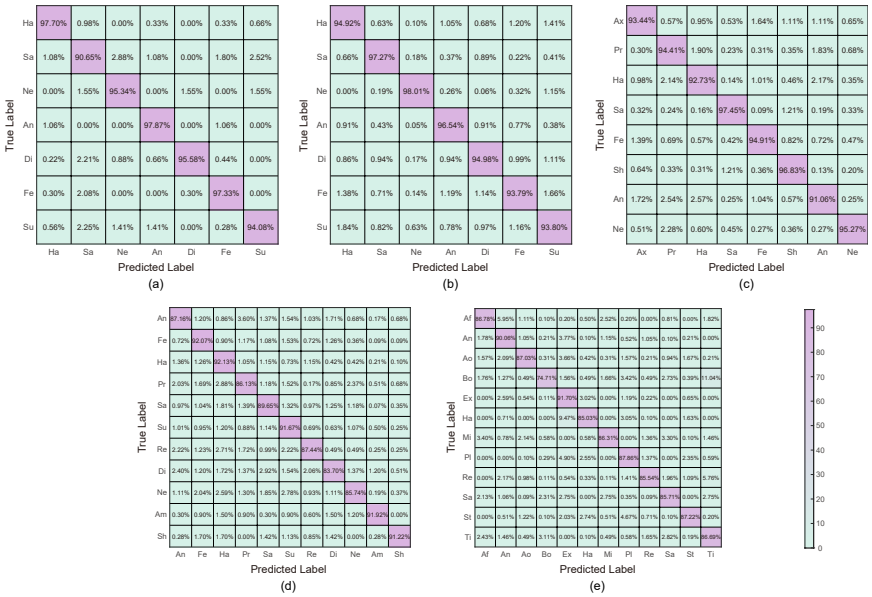

Fig. 9. The confusion matrix of emotion recognition on the (a) EGBM dataset, (b) KDAE dataset, (c) Emilya dataset, (d) MPI dataset, and (e) DMCD dataset.

TABLE 4
The performance (%) of the proposed method and other methods on the five datasets. The best results are labeled in bold.

	EGBM		KDAE		Emilya		MPI		DMCD	
Methodology	Protocol	Accuracy	Protocol	Accuracy	Protocol	Accuracy	Protocol	Accuracy	Protocol	F1-score
CNN [54]	LOSO	58.10	---	---	---	---	---	---	---	---
RNN [54]	LOSO	59.40	---	---	---	---	---	---	---	---
RNN-LSTM [54]	LOSO	69.00	---	---	---	---	---	---	---	---
AS-LSTM [55]	LOSO	74.10	---	---	---	---	---	---	---	---
MVRL [56]	---	---	10-fold	64.10	---	---	---	---	---	---
ST-GCNs [57]	---	---	10-fold	65.00	---	---	---	---	---	---
GCN [58]	---	---	---	---	---	---	5-fold	56.03	---	---
L-GrIN [58]	---	---	---	---	---	---	5-fold	58.59	---	---
RF-Motion Features [59]	---	---	---	---	3-fold	75.00	---	---	---	---
SVM-χ^2 Kernel [60]	---	---	---	---	10-fold	82.20	10-fold	78.60	---	---
Multiscale CNN [12]	---	---	---	---	10-fold	91.31	---	---	10-fold	74.68
Our method	LOSO	**83.24 ± 2.03**	---	---	3-fold	**93.01 ± 1.81**	5-fold	**88.12 ± 0.77**	---	---
	10-fold	**95.55 ± 1.47**	10-fold	**95.60 ± 0.62**	10-fold	**94.42 ± 0.68**	10-fold	**89.60 ± 1.64**	10-fold	**86.15 ± 3.02**

emotion recognition. Crenn et al. [60] extracted the posture spectral features and utilized the SVM with a χ^2 kernel to identify emotions. Beyan et al. [12] proposed a multiscale CNN structure to accommodate 8-bit RGB images obtained from full-body skeleton data for emotion recognition based on body movements.

4) For the MPI dataset, Shirian et al. [58] proposed a learnable graph inception network (L-GrIN) that jointly learned emotional representations in the underlying graph structure of the body skeleton data. In addition, the authors used a standard spectral graph convolution network (GCN) to conduct experiments.

5) For the DMCD dataset, the compared method is the multiscale CNN in [12], which allows multiple posture image formats to be used as input simultaneously.

In Table 4, we compared our approach with existing methods on the five datasets. Overall, our approach achieved better performance than recent state-of-the-art approaches on all datasets. On the EGBM, KDAE, and MPI datasets, the proposed multiscale spatio-temporal network achieved the highest recognition accuracies, which were significantly better than the classification accuracies of the CNN, LSTM, GCN and optimized networks based on L-STM and GCN architectures. This finding reveals that the proposed method is superior to conventional deep learning methods for emotion recognition from full-body motion. On the Emilya dataset, the proposed method achieved the best performance, with an average accuracy of 94.42% and a standard deviation of 0.68%, significantly outperforming the RF and SVM models based on hand-crafted posture features. This result demonstrates that our model is more suitable for emotion recognition from full-body motion than conventional machine learning classifiers. Furthermore, the average accuracy of the multiscale CNN [12] on the Emilya and DMCD datasets was 91.31% and 74.68%, respectively, which was 3.11% and 11.47% lower than that of the proposed method. This finding further indicates the effectiveness of incorporating the multiscale spatial network with Riemannian network architectures into multiscale temporal networks based on CNN architectures in this paper.

TABLE 5
The average accuracies (%) of the proposed method (PM) and other methods on eight single actions on the Emilya dataset, which included simple walking (SW), walking with an object in hands (WH), moving books on a table (MB), being seated (BS), sitting down (SD), knocking (KD), lifting (Lf), and throwing (Th). The performance of the PM for each emotion class during different actions are also given. The best results are labeled in bold.

	SW	WH	MB	BS
[59], [61]	85.00	84.00	83.00	68.00
[12]	87.29	87.35	92.02	96.59
PM	**96.69**	**96.31**	**99.35**	**97.56**
PM-Anxiety	95.25	95.70	99.45	96.94
PM-Pride	96.94	94.98	98.91	96.30
PM-Happiness	91.20	88.05	99.62	97.24
PM-Sadness	**99.65**	**99.65**	**100**	**99.04**
PM-Panic Fear	96.75	97.61	98.99	96.65
PM-Shame	97.57	97.98	99.81	98.83
PM-Anger	89.97	92.01	96.85	97.04
PM-Neutral	96.33	95.29	99.81	94.79
	SD	KD	Lf	Th
[59], [61]	68.00	82.00	78.00	79.00
[12]	**87.63**	93.03	90.24	90.10
PM	85.52	**97.05**	**97.61**	**93.29**
PM-Anxiety	79.14	97.76	97.07	95.76
PM-Pride	84.62	95.50	98.15	89.79
PM-Happiness	82.61	97.57	**98.91**	92.56
PM-Sadness	94.21	**98.61**	98.02	95.07
PM-Panic Fear	**95.50**	96.11	98.79	95.09
PM-Shame	88.17	97.99	97.24	94.28
PM-Anger	74.07	91.52	97.17	**95.83**
PM-Neutral	85.81	98.13	94.82	83.50

Finally, we evaluated the performance of the proposed method (PM) on each action class on the Emilya dataset and compared our results with the results of state-of-the-art methods presented in [12], [59] and [61]. The results are shown in Table 5. Furthermore, Table 5 shows the recognition accuracies of the proposed method for each emotion class during different actions. We followed the same cross validation settings as in the above comparative literature. As shown in Table 5, in the evaluation of individual action classes, the PM outperforms the existing methods, with performance improvements of 0.97-29.56% for 7 out of 8 actions. With the exception of the SD action, the recognition accuracies of the PM exceed 93% for 7 actions, and the PM achieves the highest classification results for the MB action, with an accuracy of 99.35%. In addition, it can be observed that sadness was recognized with the highest classification accuracy for 5 out of 8 actions, while the lowest accuracy was obtained for anger, which is similar to the results in [12].

6 Conclusion and Future Work

In this paper, we propose a multiscale spatio-temporal network for emotion recognition based on full-body motion. First, we innovatively design an adaptive scale selection algorithm based on the pseudo-energy model, which guides our network to focus on long-term macroscopic body expression (coarse-grained modeling) and short-term subtle emotional posture changes (fine-grained modeling). Then, we construct a hierarchical network architecture based on the Riemannian network and CNN, which can jointly extract the spatio-temporal affective representations encoded in the posture covariance matrices and 3D posture images with different time scales. Finally, a ST-ITE fusion algorithm is proposed, which enables the network to perform effective spatio-temporal optimization while reducing overfitting during network fusion. The experimental results on five public datasets indicate that multiscale analysis of full-body motion can provide more discriminative emotion representations. The proposed ST-ITE fusion algorithm improves the classification performance and significantly enhances the generalizability and convergence of the model. Furthermore, the proposed method achieves excellent results on limited data and data with class imbalances, and outperforms state-of-the-art methods on all datasets. These results demonstrate the superiority and robustness of the proposed multiscale spatio-temporal network for emotion recognition based on full-body motion.

However, there are several limitations of this research that need to be addressed in future work. First, the proposed energy-based scale selection algorithm may struggle to capture the fine-grained features in high complexity datasets (e.g. DMCD dataset). This limitation may arise from the simplified nature of the pseudo-energy model. For future work, we plan to incorporate more advanced synthesis method of neutral posture [60] into the proposed pseudo-energy model. Furthermore, we also plan to assign weights to the different energy models in the pseudo-energy model to better combine the complementarity of the kinetic and potential energy. These methods will further improve the ability of the scale selection method in capturing the richness and subtlety of complex body expressions.

Second, the proposed multiscale spatio-temporal network ignores the varying contributions of different skeletal joints to emotion recognition, resulting in the inclusion of redundant joints that may affect recognition performance. To address this limitation, future work could introduce an attention mechanism into the network, which enable the network to focus on joints that contain abundant emotional information while suppressing the influence of redundant or less informative joints. In addition, we plan to introduce data augmentation methods into the network to further improve the robustness of the model.

In the future, we will also further explore the potential application of the proposed emotion recognition model based on full-body motion in the field of mental health. Specifically, our research can analyse the emotional state of multiple individuals simultaneously in a relatively short period of time, which will provide new methods and insights for large-scale early detection of mental disorders. This will further advance the development of low-cost, non-invasive, and intelligent systems for the diagnosis and treatment of mental disorders. Furthermore, the proposed emotion recognition model is valuable in HCI. Traditional facial expression recognition has limitations when facial expressions are hindered, such as wearing masks. In contrast, our approach can capture a wider range of emotional information by analyzing full-body movements, thus overcoming the limitations of traditional facial expression recognition and providing a more comprehensive and accurate emotion recognition result. In addition, the proposed emotion recognition method has great potential in various fields such as education, virtual reality, gaming, and social robotics.

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.

REFERENCES

[1] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 406–417, 2014.

[2] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, 2018.

[3] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[4] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Mach. Vision Appl.*, vol. 30, no. 5, pp. 975–985, 2019.

[5] Q. Ma, L. Shen, E. Chen, S. Tian, J. Wang, and G. W. Cottrell, "Walking walking walking: Action recognition from action echoes," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2457–2463.

[6] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *J. Visual Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, 2014.

[7] J. Fang, T. Wang, C. Li, X. Hu, E. Ngai, B.-C. Seet, J. Cheng, Y. Guo, and X. Jiang, "Depression prevalence in postgraduate students and its association with gait abnormality," *IEEE Access*, vol. 7, pp. 174 425–174 437, 2019.

[8] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[9] T. Wang, C. Li, C. Wu, C. Zhao, J. Sun, H. Peng, X. Hu, and B. Hu, "A gait assessment framework for depression detection using kinect sensors," *IEEE Sens. J.*, vol. 21, no. 3, pp. 3260–3270, 2020.

[10] European FET PROACTIVE Project EnTimeMent. [Online]. Available: http://entimement.dibris.unige.it.

[11] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, "The dancer in the eye: towards a multilayered computational framework of qualities in movement," in *Proc. 3rd Int. Symp. Movement Comput.*, 2016, pp. 1–7.

[12] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, and R. Niewiadomski, "Modeling multiple temporal scales of full-body movements for emotion classification," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1070–1081, 2023.

[13] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, and A. Del Bimbo, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 550–560.

[14] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 1, pp. 1–31, 2016.

[15] J. Shi, C. Liu, C. T. Ishi, and H. Ishiguro, "Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network," *Sensors*, vol. 21, no. 1, 2020, Art. no. 205.

[16] S. Karumuri, R. Niewiadomski, G. Volpe, and A. Camurri, "From motions to emotions: classification of affect from dance movements using deep learning," in *Proc. Extended Abstracts CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–6.

[17] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2036–2042.

[18] Y. LeCun, Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 255–258, 1995.

[19] J. Kong, Y. Bian, and M. Jiang, "Mtt: Multi-scale temporal transformer for skeleton-based action recognition," *IEEE Signal Process Lett.*, vol. 29, pp. 528–532, 2022.

[20] W. Li, X. Liu, Z. Liu, F. Du, and Q. Zou, "Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network," *IEEE Access*, vol. 8, pp. 144 529–144 542, 2020.

[21] J. Wang, Y. Lin, M. Zhang, Y. Gao, and A. J. Ma, "Multi-level temporal dilated dense prediction for action recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 2553–2566, 2022.

[22] F. Cheng, H. Zheng, and Z. Liu, "From coarse to fine: Hierarchical multi-scale temporal information modeling via sub-group convolution for video action recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[23] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–6.

[24] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 106–118, 2011.

[25] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on microsoft kinects," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 585–591, 2018.

[26] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Proc. IEEE Conf. Syst., Man Cybern.*, 2005, pp. 3437–3443.

[27] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, and J. C. Alvarez-Paiva, "A novel geometric framework on gram matrix trajectories for human behavior understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 1–14, 2020.

[28] J. Gao, Y. Guo, and Z. Wang, "Matrix neural networks," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 313–320.

[29] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Matrix backpropagation for deep networks with structured layers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2965–2973.

[30] Y. Ollivier, "Riemannian metrics for neural networks I: Feedforward networks," *Inf. Inference J. IMA*, vol. 4, no. 2, pp. 108–153, 2015.

[31] R. Wang, X. J. Wu, T. Xu, C. Hu, and J. Kittler, "U-SPDNet: An spd manifold learning-based neural network for visual classification," *Neural Networks*, vol. 161, pp. 382–396, 2023.

[32] Z. Ding, P. Wang, P. O. Ogunbona, and W. Li, "Investigation of different skeleton features for cnn-based 3d action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2017, pp. 617–622.

[33] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. IEEE 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.

[34] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4570–4579.

[35] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2842–2855, 2018.

[36] S. Laraba, M. Brahimi, J. Tilmanne, and T. Dutoit, "3D skeleton-based action recognition by representing motion capture sequences as 2D-RGB images," *Comput. Anim. Virtual Worlds*, vol. 28, no. 3/4, 2017, Art. no. e1782.

[37] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *Proc. Int. Conf. Pattern Recognit. Workshops*, 2019, pp. 153–163.

[38] M. Zhang, L. Yu, K. Zhang *et al.*, "Kinematic dataset of actors expressing emotions," *Scientific data*, vol. 7, no. 1, pp. 1–8, 2020.

[39] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 90–101, 2018.

[40] E. Volkova, S. De La Rosa, H. H. Bülthoff, and B. Mohler, "The MPI emotional body expressions database for narrative scenarios," *PLoS ONE*, vol. 9, no. 12, 2014, Art. no. e113647.

[41] Dance Motion Capture Database (DMCD). [Online]. Available: http://dancedb.eu/.

[42] H. Lu, S. Xu, X. Hu, E. Ngai, Y. Guo, W. Wang, and B. Hu, "Postgraduate student depression assessment by multimedia gait analysis," *IEEE MultiMedia*, vol. 29, no. 2, pp. 56–65, 2022.

[43] M. A. Mahfoudi, A. Meyer, T. Gaudin, A. Buendia, and S. Bouakaz, "Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and valida-

tion," *IEEE Trans. Affect. Comput.*, early access, 2022. [Online]. Available: https://doi.org/10.1109/TAFFC.2022.3226252.
[44] C. Tang, W. Li, P. Wang, and L. Wang, "Online human action recognition based on incremental learning of weighted covariance descriptors," *Inf. Sci.*, vol. 467, pp. 219–237, 2018.
[45] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on riemannian manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[46] X. Zhang, D. Lu, J. Shen, J. Gao, X. Huang, and M. Wu, "Spatial-temporal joint optimization network on covariance manifolds of electroencephalography for fatigue detection," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2020, pp. 893–900.
[47] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, 2007.
[48] M. T. Wu, "Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022.
[49] S. G. Zadeh and M. Schmid, "Bias in cross-entropy-based training of deep survival networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3126–3137, 2020.
[50] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.
[51] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 140–149.
[52] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vol. 208–209, 2021, Art. no. 103219.
[53] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, 2015.
[54] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari, "Emotion recognition from skeletal movements," *Entropy*, vol. 21, no. 7, pp. 646–661, 2019.
[55] H. Zhang, P. Yi, R. Liu, and D. Zhou, "Emotion recognition from body movements with AS-LSTM," in *Proc. IEEE Int. Conf. Virtual Rehabil.*, 2021, pp. 26–32.
[56] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and G. L. Foresti, "Affective action and interaction recognition by multi-view representation learning from handcrafted low-level skeleton features," *Int. J. Neural Syst.*, vol. 32, no. 10, 2022, Art. no. 2250040.
[57] E. Ghaleb, A. Mertens, S. Asteriadis, and G. Weiss, "Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2021, pp. 1–8.
[58] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Trans. Multimedia*, vol. 24, pp. 780–790, 2022.
[59] N. Fourati, "Classification and characterization of emotional body expression in daily actions," Ph.D. dissertation, Télécom ParisTech, France, 2015. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01282785.
[60] A. Crenn, A. Meyer, H. Konik, R. A. Khan, and S. Bouakaz, "Generic body expression recognition based on synthesis of realistic neutral motion," *IEEE Access*, vol. 8, pp. 207 758–207 767, 2020.
[61] N. Fourati, C. Pelachaud, and P. Darmon, "Contribution of temporal and multi-level body cues to emotion classification," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 116–122.

Tao Wang received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2018, and the M.S. degree from Lanzhou University, Lanzhou, China, in 2021. He is currently pursuing the Ph.D. degree in the Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China. His research focuses on affective computing, bioelectrical signal processing, and automated mental disorder detection.

Shuang Liu received the B.S. degree in biomedical engineering from Tianjin Medical University, Tianjin, China, in 2012, and the M.S. and Ph.D. degrees in biomedical engineering from Tianjin University, Tianjin, in 2018. She is currently an Associate Professor with the Academy of Medical Engineering and Translational Medicine, Tianjin University. Her research interests include physiological mechanism of emotion, emotion recognition and regulation, and biomarker detection of the depression.

Feng He received the B.S. and M.S. degrees in biomedical engineering from Tianjin University, Tianjin, China, in 1994 and 1998, respectively. Since 2015, he has been a Professor with the College of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin, China. His research interests include neural engineering, biomedical signal detection and processing, and medical instrument design.

Weina Dai received the B.Eng. degree in College of Medicine and Biological Information Engineering from the Northeastern University, Shenyang, China, in 2021. She is currently pursuing the M.S. degree in School of Precision Instrument and Opto-electronics Engineering from the Tianjin University, Tianjin, China. Her research interests are in emotion recognition, computer vision, and machine learning.

Minghao Du received the B.S. degree in communication engineering from Wuhan University of Technology, Wuhan, China, in 2020. He is currently pursuing the M.S. degree with the Academy of Medical Engineering and Translational Medicine from Tianjin University, Tianjin, China. His research interest focuses on affective computing, deep learning, speech analysis, depression estimation and classification.

Yufeng Ke completed the Ph.D. degree from the Department of Biomedical Engineering, College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin, China, where he is currently an Associate Professor with the Academy of Medical Engineering and Translational Medicine. His research interests include brain-computer interfaces, brain stimulation and adaptive human-machine interaction systems.

Dong Ming received the B. S. and Ph.D. degrees in biomedical engineering with Tianjin University, Tianjin, China, in 1999 and 2004, respectively. During 2005–2006, he was a Visiting Scholar with the Division of Mechanical Engineering and Mechatronics, University of Dundee, Dundee, U.K. In 2006, he joined Tianjin University (TJU) Faculty, College of Precision Instruments and Optoelectronics Engineering and since 2011, has been a Full Professor of biomedical engineering. He is currently the Dean of the Academy of Medical Engineering and Translational Medicine of TJU, the Head of the Neural Engineering and Rehabilitation Laboratory, TJU, and the Chair of IEEE EMBS Tianjin Chapter. His main research interests include neural engineering, rehabilitation engineering, sports science, biomedical instrumentation and signal/image processing, especially in functional electrical stimulation, gait analysis, and brainCcomputer interface.