Received: 9 November 2018 | Revised: 11 March 2019 | Accepted: 23 March 2019

DOI: 10.1111/coin.12207

SPECIAL ISSUE ARTICLE

WILEY Computational Intelligence

Deep spatiotemporal LSTM network with temporal pattern feature for 3D human action recognition

Yirui Wu[1,2] | Lianglei Wei[3] | Yucong Duan[4]

[1]College of Computer and Information, Hohai University, Nanjing, China

[2]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

[3]Ruiting Network Technology Company, Shanghai, China

[4]College of Information and Technology, Hainan University, Haikou, China

Correspondence
Yucong Duan, College of Information and Technology, Hainan University, 58 People Avenue, Haikou 570228, China.
Email: duanyucong@hotmail.com

Funding information
National Key R&D Program of China, Grant/Award Number: 2018YFC0407901; National Natural Science Foundation of China, Grant/Award Number: 61702160, 61872219, and 61702277; Fundamental Research Funds for the Central Universities, Grant/Award Number: 2016B14114; Natural Science Foundation of Jiangsu Province, Grant/Award Number: BK20170892; National Key Laboratory for Novel Software Technology, NJU, Grant/Award Number: K-FKT2017B05

Abstract

With the rapid development of RGB-D cameras and pose estimation techniques, action recognition based on three-dimensional skeleton data has gained significant attention in the artificial intelligence community. In this paper, we incorporate temporal pattern descriptors of joint positions with the currently popular long short-term memory (LSTM)–based learning scheme to obtain accurate and robust action recognition. Considering that actions are essentially formed by small subactions, we first utilize a two-dimensional wavelet transform to extract temporal pattern descriptors in the frequency domain for each subaction. Afterward, we design a novel LSTM structure to extract deep features, which model a long-term spatiotemporal correlation between body parts. Since temporal pattern descriptors and LSTM deep features can be regarded as multimodal representations for actions, we fuse them with an autoencoder network to achieve a more effective feature descriptor for action recognition. Experimental results on three challenging data sets with several comparative methods demonstrate the effectiveness of the proposed method for three-dimensional action recognition.

KEYWORDS

long short-term memory, spatiotemporal analysis, video analysis, 3D action recognition

1 | INTRODUCTION

Nowadays, how to build smart environments for enhancing the quality of life has drawn increasing attention from researchers.[1-3] Among the topics in smart environments, one popular issue is

understanding the meanings of users' actions, in order to decide how to react properly to users' behavior.[4-6] Recently, videos with three-dimensional (3D) skeleton data could be easily achieved with the rapid development of low-price RGB-D camera, ie, Kinect and Intel RealSense, which results in making 3D human action recognition a hot and new challenge in content-oriented video analysis.[7,8] Various methods for feature extraction and classifier learning thus have been developed for 3D human action recognition.[9-11]

We generally group current 3D human action recognition methods into two categories, namely, hand-crafted feature–based methods[5,9,10] and deep neural network methods.[11,12,13] Hand-crafted feature–based methods design various kinds of features, such as histogram of oriented gradients (HOG),[14] Cuboids,[15] extended SURF,[16] and so on, to visually and temporally describe human motion sequences. For example, Wang et al[17] proposed the conception of dense trajectories to describe action sequences by a number of hand-crafted features focused on both motion and appearance, which achieves desirable results on multiple data sets. Generally speaking, hand-crafted features are clear in design purpose and easy to be interpreted as two-stage methods of feature extraction and classification. However, such methods usually fail in discovering hidden patterns from the quantity of 3D skeleton data, which can be beneficial for more accurate and robust action recognition.

Another category, ie, deep neural network methods, learns spatiotemporal characteristics by automatically extracting distinctive features from large data for accurate recognition.[18,19] Among the different neural-based architectures, recurrent neural networks (RNNs), which are specially designed to handle sequential data with variable length, have achieved promising performances in 3D action recognition.[20,21] For example, Liu et al[13] proposed a long short-term memory (LSTM) network incorporating a tree structure to describe the relation of human parts, which successfully utilizes the spatiotemporal characteristics of human actions for the recognition task and achieves desirable accuracy on a large data set, ie, NTU RGB+D.[22] Following on the thought of the modeling relationship of two concurrent domains, ie, spatial and temporal, Hu et al[23] proposed a deep bilinear framework to further describe such relationship, where their proposed modality pooling layer and temporal pooling layer could pool the input action sequence along the modality and temporal directions separately. By the different means of describing spatiotemporal characteristics of human actions, hidden patterns of human actions are more clearly interpreted by researchers to increase recognition accuracy. However, human actions are complicated to describe and variant in patterns from one person to another. It is hard to guarantee rationality and robustness by only utilizing deep features without considering expert knowledge from hand-crafted features.

In fact, utilizing an RGB-D camera for action recognition suffers from not only variances of patterns of actors but also the fact that recognized skeletons cannot always be accurate due to illumination variations, noise, occlusions, and so on. Therefore, it is not applicable for most of the previous deep neural–based methods to deal with videos captured in real-life scenarios,[24] since they require reliable RGB-D input streams and only utilize information captured from a relatively small number of people for training. To solve such a problem, we thus propose to couple the strength of hand-crafted feature–based methods and deep neural network methods. The proposed method essentially originates from the thought that expert knowledge behind hand-crafted features could focus on the informative parts of action sequence and help utilize a small data set for effective recognition, which is a beneficial complement to deep neural network methods facing difficulties of pattern variances and noisy input. In other words, the fusion of heterogeneous features, ie, hand-crafted and deep neural features, could improve robustness of action recognition by analyzing the action sequence from different aspects, ie, an expert view and a data-driven model view, respectively.

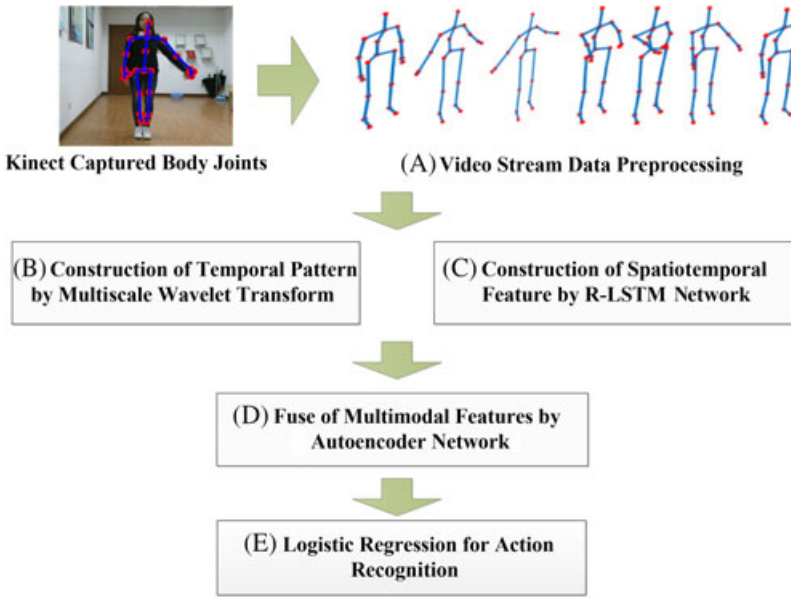

FIGURE 1 Workflow of the proposed method for three-dimensional action recognition, which consists of (A) preprocessing on the inputting of raw body joint data captured by Kinect, (B) utilizing a multiscale wavelet transform to extract temporal patterns from subactions, (C) using the relation-aware long short-term memory (R-LSTM) network to extract spatiotemporal features, (D) fusion of extracted multimodal features with an autoencoder network, and (E) action recognition by performing classification on the fused features [Color figure can be viewed at wileyonlinelibrary.com]

With the idea of utilizing the advantages of hand-crafted feature–based methods and deep neural network methods, we design the workflow of the proposed method, as shown in Figure 1. During Step (A), the proposed method performs data preprocessing on the captured raw 3D data of the human body, which not only transforms and rescales joint point positions based on camera view variations but also extracts key action frames to relieve the burden of high computation and improve robustness in terms of noise input. Based on the stable preprocessed data, Step (B) of the proposed method transforms subactions, ie, meaningful parts of the action sequence, into the frequency domain and extracts the temporal features of subactions by a multiscale wavelet transform. Since subactions can be regarded as local states of the action sequence, we thus describe the action sequence in a local sense by representing action as subactions and learning relationship between subactions. Meanwhile, Step (C) of the proposed method extracts deep spatiotemporal features from the whole action sequence based on a novel LSTM network, which successfully captures the relationship between human body parts in a relatively global sense. During Step (D), the proposed method utilizes an autoencoder (AE)-based fusion network to not only model the shared and informative components of hand-crafted and deep features but also involve their specific features with discriminative information for a better action representation, which is further adopted for recognition purposes with a logistic regression method in Step (E).

There are three major contributions of this paper.

- Introduction of temporal pattern descriptors in the time-frequency domain extracted by a multiscale wavelet transform, which describes the action sequence via local subactions and is invariant to multiple unpleasant impacts, such as noise, temporal misalignment, and the translation of the human body.

- Introduction of spatiotemporal features extracted from the proposed relation-aware long short-term memory (R-LSTM) network, which discovers hidden patterns of actions from the quantity of training samples and emphasizes the global modeling of joints relation factors.
- A highly efficient fusing method is introduced to fuse hand-crafted temporal pattern features and deep neural features. The experimental results on several public data sets show that fusion features outperform individual features on action recognition performance.

2 | RELATED WORK

In the following subsections, we first review the approaches for 3D action recognition based on the categories of hand-crafted feature–based methods and deep neural network methods. Afterward, we introduce the most recent process of 3D action recognition by involving multimodal data, which is closely related to our work.

2.1 | Hand-crafted feature–based methods for 3D action recognition

Researchers have carefully observed the characteristics of human actions from different aspects. Based on such observation, feature extraction for different purposes, such as descriptors on skeleton geometry and joint dynamic or appearance information, and associated classifier learning approaches are utilized for 3D action recognition.[25,26] We introduce several most popular and related feature designs in the following parts, namely, skeleton geometry and joint dynamic information.

Skeleton geometric information depicted in depth sequences can be used to characterize action. For instance, Oreifej and Liu[27] utilized the histograms of oriented normal within each spatiotemporal depth cube to describe actions. However, a simple histogram representation cannot provide sufficient distinguishing ability for complicated skeleton patterns. Later, Evangelidis et al[28] studied skeletal quads to assist in action recognition, which is actually a learned Gaussian mixture model distribution over the Fisher kernel representation, which acts as a succinct skeletal feature for action recognition and achieves remarkable accuracy. Meanwhile, Vemulapalli et al[10] first represented skeleton configurations and actions as points and curves in a Lie group and then utilized a support vector machine classifier to classify action categories, which offers an optional framework to describe skeleton based on graph knowledge. Skeletal joints are also popularly modeled as a tree-based pictorial structure.[13,29] It is beneficial to model the spatial dependency of the joints based on their adjacency tree structure, since a hidden representation of several joints such as the neck joint could be more informative than that of other joints, such as the right- or left-hand joints, which is generally the same as expected. It is noted that we follow such tree representation to describe skeleton geometry due to its effectiveness and simplicity.

Human action can also be characterized by the dynamics of human poses utilizing time domain analysis.[30-32] Early on, Xia et al[33] proposed a hidden Markov model–based method to model the temporal dynamics of the actions over a histogram-based representation of 3D joint locations. Still, a simple histogram representation cannot provide sufficient distinguishing ability for dynamics of human poses. Later, an angular skeletal representation over the tree-structured set of joints is then introduced in the work of Ohn-Bar and Trivedi,[34] which calculates the similarity of these features over temporal dimension to build the global representation of the action samples. However, actions with the same meaning can be formed by different combinations of subactions, with its simplicity of calculating similarity in the time domain, thus preventing its

further usage in real-life applications. Later, Liang et al[35] first extracted the time-domain features of action sequences through hierarchical depth motion maps (LDM) and then extracted the spatial domain features by a multiscale direction gradient histogram (HOG) operator. After feature extraction, they utilized an improved sparse coding method to combine the extracted features for classification of actions, which provides a simple but effective idea of involving both time- and spatial-domain descriptors. Most recently, Fernando et al[36] used a function-based temporal pooling method to capture the latent structure of the video sequence data and how frame-level features evolve over time in a video. In comparison, our method explores the collaboration and naturally fusing among geometric and temporal information, and thus, the weakness of using single information can be overcome by working collaboratively.

2.2 | Deep neural network methods for 3D action recognition

Since the recent resurgence of neural networks invoked by Hinton et al,[37] deep neural networks have become an effective approach to extracting high-level features from massive data. Following this trend, researchers from the artificial intelligence community have tried different categories of deep neural models to purse a more accurate 3D action recognition.[12,38,39] Since RNNs could handle sequential data with variable length coinciding with features of action sequences of multiple persons, we pay special attention on such architecture and its modifications, ie, LSTMs, in the following review.

Early on, a hierarchical bidirectional RNN[40] applies bidirectional RNNs in a novel hierarchical fashion, such that they divide the entire skeleton into five major groups of joints and that each group was fed into a separated bidirectional RNN. Because of the disadvantage of RNN-based methods, ie, vanishing gradient problem, a special kind of RNN named LSTM has been popular in human action recognition, which utilizes a gating mechanism over an internal memory cell to learn and retain both long- and short-term dependencies in sequential input data. For example, Veeriah et al[41] proposed a differential gating scheme for the LSTM neural network, which emphasizes on the change in information gain caused by the salient motions between the successive frames, which is similar in thought with the design of the proposed LSTM architecture. Later, Shahroudy et al[22] separated the memory cell to part-based subcells and pushed the network toward learning the long-term context representations individually for each part, which offers a novel idea on utilizing part-level time-domain information for action recognition. Focusing on spatiotemporal information, ST-LSTM[13] explores spatiotemporal domains to analyze the hidden sources of action-related information within the input data over both domains concurrently, which represents the topology of the human body as a traversal tree structure and proposes a novel trust gate to improve accuracy in terms of noisy input. Emphasizing on the co-occurrence property of joints, Zhu et al[12] proposed a mixed-norm regularization term to a deep LSTM network's cost function, which successfully pushes the network toward learning the co-occurrence of discriminative joints for action classification.

Most recently, several LSTM-based methods have been using different streams to perform action recognition for higher accuracy and robustness. Wang and Wang[42] proposed a novel two-stream RNN architecture to model both temporal dynamics and spatial configurations for skeleton-based action recognition, which provides a novel idea by utilizing two different architectures for the description of temporal and spatial information. Moreover, the attention model,[43,44] ie, the selectively focusing mechanism, is popular in 3D action recognition, in order to focus on informative parts of joints or key frames. Song et al[45] built their attention-based model on top

of the LSTM architecture, which learns to selectively focus on discriminative joints of skeleton within each frame of the inputs and pays different levels of attention to the outputs of different frames. Furthermore, Liu et al[46] proposed a global context-aware attention LSTM for RGB-D action recognition, which recurrently optimizes the global contextual information and further utilizes it as an informative function to assist in accurate action recognition. It is noted that the high accuracy of deep neural network methods results in their wide usage in real-life applications, which are specially designed for massive usage with technologies, such as cloud computing,[47,48] edge computing,[49,50] big data technology,[51] and so on. Inspired by the work of Veeriah et al,[41] here, the proposed method takes the difference between the current frame and the previous one as the input value to reduce the impact of body parts and thus tries to model differential relationship calculation of each joint part between frames by LSTM to improve accuracy.

2.3 | Multimodal 3D action recognition

We outline the methods that learn multimodal features for action recognition, since integrating multimodal features can generally improve the recognition performance. An intuitive way to combine multimodal features is to directly concatenate them together.[52] To mine more useful information among multimodal features for better performance, researchers propose to explicitly learn shared-specific structures among features.[11,53]

Early on, Liu and Shao[54] utilized a genetic programming framework to improve not only RGB and depth descriptors but also their fusion simultaneously through an iterative evolution. Ni et al[55] concatenated depth descriptor– and RGB-based representations of spatiotemporal interest points for better RGB+D information fusion. Then, Song et al[56] achieved accurate RGB+D action recognition by tracking trajectories consisted of interest points and describing these points via depth-base local surface patches. The work of Kong and Fu[57] first applies projection matrices to the independent spaces between RGB and depth modalities and then learns models by minimizing the rank with their proposed low-rank bilinear classifier. Most recently, Shahroudy et al[11] proposed a new deep AE-based shared-specific feature factorization network to separate input multimodal signals into a hierarchy of components. Similar to the aforementioned work,[11] we apply an AE as a highly efficient fusing method to fuse hand-crafted temporal pattern features and deep neural features for better robustness and accuracy.

3 | VIDEO STREAM DATA PREPROCESSING

In this section, we first perform filter calculation and normalization to deal with the misalignment, scaling, and view changing problems of an input raw stream, respectively. Afterward, we extract key frames as input for later steps for the benefits of stability and less computing complexity.

We utilize Kinect v2.0 to capture body actions, which tracks 25 body joints, and each joint i has 3D coordinates $j_i^t = [x_i^t, y_i^t, z_i^t]$ at time t. Skeleton data acquired by Kinect have the advantages of small size and well-structured data, but suffer from drawbacks such as instability and noise in some situations. We represent some usual cases of disadvantages in Figure 2, where image (A) refers to the ideal captured skeleton image, image (B) represents misalignments of body joints caused by occlusions or the exceeding out of the sensor range, image (C) happens when the distance between the camera and the body gets larger, and image (D) is caused by view changes of the camera. It is noted that body joint misalignments are the main noise source of captured sequences;

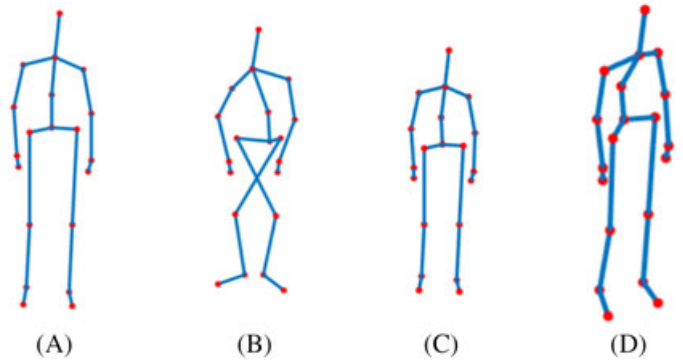

FIGURE 2 Several usual undesirable cases of Kinect-captured raw skeleton data. A, The ideal captured image; B, Misalignment cases of joints; C, Scaling problem of body joints; D, Side view cases of the skeleton image [Color figure can be viewed at wileyonlinelibrary.com]

meanwhile, scaling and view changes of cameras could enlarge the intraclass dissimilarity of body actions, thereby increasing the difficulty of action recognition.

To solve the problem of misalignments, we thus apply two filters on the captured raw data to stabilize sculptor actions.

1. **Holt-Winters double exponential smoothing filter**. This filter applies Holt-Winters double exponential smoothing to historical joint positions and orientations to get predictions on upcoming action data under a reasonable assumption that there exists a trend in the captured action data. By interpolating between the prediction and upcoming action data, this filter stabilizes joint locations and orientations to remove most jitters and noise brought by Kinect.
2. **Limbs inferring filter**. This filter is applied to the positions of occlusive limbs, aiming at preventing the jumpy of limbs. Kinect can provide rough predictions for clipped limb joints; however, the inference can occasionally be erroneous since it is based on a limited-depth image. We thus linearly interpolate the previous smoothed joint positions and the inferred positions to predict convinced positions for clipped limbs.

After stabilizing the raw input stream of Kinect-captured skeleton data, we normalize the coordinates of joints to make the input stream invariant for orientation or scale invariants, which could help solve the problem of scaling and view changes to a certain extent. Specifically, we first fit the left shoulder and torso joints to the x-axis and then calculate the rotation angle during fitting. After that, we fit bones between the head and torso joints to the xy-plane, where the rotation angle is calculated and used to rotate rest bones. After such operations, we could achieve a stable and less noisy input skeleton stream.

When humans try to recognize actions, we observe that they generally classify actions based on information from key frames. Based on such observations, we believe that frames of action sequence are redundant to achieve action recognition results. In other words, information from key frames is sufficient to guarantee the accuracy and robustness of action recognition. Essentially, key frame extraction could reduce the intern-class difference to decrease the difficulty in recognizing the same category action performed by different persons. Moreover, key frame extraction not only helps eliminate the noise of action sequence, such as joint misalignment cases during several intern frames, but also largely reduces the input size and, thus, decreases computation cost during processing.

Following the general idea for key frame extraction of first merging frames into clusters and then choosing the key frame from each cluster, we use the agglomerative hierarchical clustering algorithm to perform key frame extraction. Specifically, agglomerative hierarchical clustering first initializes small clusters with a single frame and then calculates similarity between adjacent clusters. Two adjacent clusters with the highest similarity are merged. This merging process is repeated until clusters with a predefined key frame number are achieved. Considering the inherent temporal characteristics of an action sequence, we utilize the dynamic time warping (DTW)[58] distance for similarity calculation. After key frame extraction, we could achieve a set of positions of body joints $J = \{j_i^t | t = 1, \ldots, n_k, i = 1, \ldots, 25\}$, where n_k is defined as the number of key frames.

4 | CONSTRUCTION OF TEMPORAL PATTERN BY MULTISCALE WAVELET TRANSFORM

This section gives a detailed description of our proposed temporal pattern feature, named wavelet temporal pattern (WTP), which is extracted by a multiscale wavelet transform. It is true that human actions have specific temporal structures.[59] In other words, one action may contain several consecutive subactions. For example, the "drink water" action may consist of two subactions, namely, "raise the cup" and "drink." By modeling the temporal relationship of subactions, we can distinguish between similar actions. Based on this idea, we propose to adaptively divide each action into combinations of subactions by DTW-based hierarchy clustering at first and then utilizing a two-dimensional wavelet transform to extract patterns of subactions in the time-frequency domain. The whole process of WTP construction is shown in Figure 3.

Different from the work of Wang et al,[59] which adopts pyramids to mechanically divide each action into subactions, we suppose that actions can be represented as short-time sequences

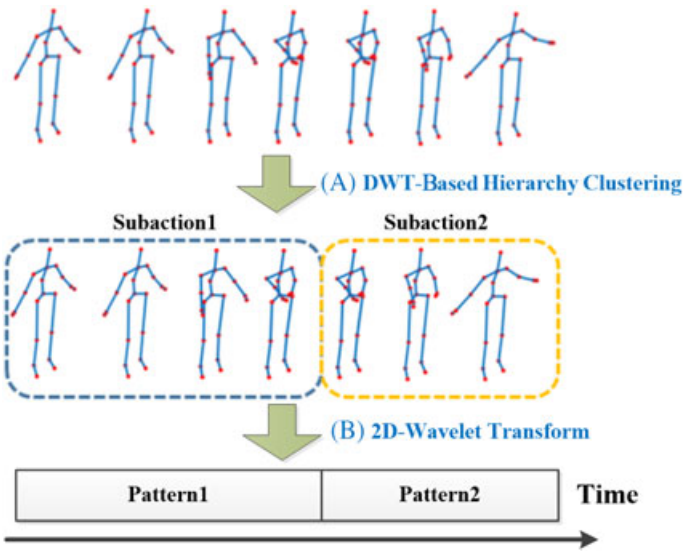

FIGURE 3 Illustration of temporal pattern feature (wavelet temporal pattern [WTP]) construction, where (A) represents dynamic time warping (DTW)–based hierarchy clustering to divide actions into subactions, and (B) refers to a multiscale two-dimensional (2D) wavelet transform, which results in WTP features to describe the temporal characteristics of extracted subactions [Color figure can be viewed at wileyonlinelibrary.com]

formed by key frames. In other words, we adopt key frames as the basic components of action. Furthermore, we suppose subactions as clusters of key frames that are near in distance. This is true for many types of actions, such as "drink water" and "pick up," where the latter could be represented as "bend the body" and "pick." Although the constructed subactions share less semantic meanings, we still argue that the components near in distance could be regarded as functional parts to represent the inherent meanings of actions. Therefore, we model the temporal relationship of actions with subactions. Based on such hypothesis, we utilize hierarchy clustering to iteratively aggregate the components, ie, key frames, to form subactions. Since the actions are temporal trajectories of body joints, we use DTW to calculate the distance between two components. Any two nearby components that own the lowest distance will be emerged so that we can construct a cluster tree from bottom to top. We adaptively decide on the number of clusters n_s (the number of subactions) by maximizing the silhouette value and a preset upper bound of n_s. We thus get a set of disjoint subactions $S = \{s_j \,|\, j = 1, \ldots, n_s \wedge s_j \in J\}$.

Regarding a subaction as a signal where joint positions vary with time, the wavelet transform helps transform the subaction into a time-frequency domain with different scales. We thus apply a two-dimensional wavelet transform, represented as $\varphi()$, to extract the low-frequency pattern of subactions, with scales varying from 1 to n_l, where n_l represents the total level number. In other words, we will abandon the high-frequency coefficients part for levels 1 to n_l during transform. We adopt the low-frequency parts as temporal patterns for subactions due to the fact that the low-frequency part is often the fundamental part for the temporal sequence. After extracting, we concatenate the transformed patterns in all scales to form the temporal pattern feature, ie,

$$F_w = \left[\varphi_1(s_j), \ldots, \varphi_{n_l}(s_j) \,|\, j = 1, \ldots, n_s\right]. \tag{1}$$

Note that each level of wavelet transform adopts the strategy of half downsampling on results computed by the last level. In other words, the size of levels decreases in half for all subactions. Since the action is set to the determined size n_k, the size of F_w will be determined as $(n_k + 1/2 \cdot n_k + \cdots + (1/2)^{n_l} \cdot n_k)$.

5 | CONSTRUCTION OF SPATIOTEMPORAL DEEP FEATURE

In this section, we aim to extract spatiotemporal deep features for action recognition based on the proposed LSTM structure, which is named R-LSTM. In other words, we aim to extract features from R-LSTM, which is designed based on differential thought between successive frames[60] and trained as a multilabel classifier to assign category labels for action sequences in the training set.

Recall that a typical LSTM unit consists of an input gate i, a forget gate f, an input modulation gate g, an output gate o, an output state h, and an internal memory cell state c. By utilizing a gating mechanism, the LSTM unit could learn and memorize a complex representation for long-term dependencies at memory cell c among the input sequence data. More detailed, the representation in c is constructed as a combination of former memory information after forgetting and new information generated from input, ie, $c^t = f^t \odot c^{t-1} + i^t \odot g^t$ at time t, where $\odot$ denotes element-wise multiplication.

Instead of keeping the long-term memory of the entire body's motion in the cell, Shahroudy et al[22] proposed a part-aware LSTM model, which keeps the context of each body part independently. In this way, the output gate will be determined by the memory of body parts instead. The idea of keeping the memory on body parts is intuitive due to the fact that body joints move together in groups.[22] This thus divides the body into five body parts, ie, body, left hand, right hand,

left leg, and right leg. The modeling of interaction between body parts in the work of Shahroudy et al[22] could help improve recognition by first involving geometrical characteristics between body parts and then modeling spatiotemporal relation[61] with the help of the LSTM structure. We keep the same idea with Shahroudy et al.[22] Moreover, it is a fact that not all body parts are useful for action recognition, since some of the body parts change little during an action. Inspired by this fact, we involve differential values of the same body part between successive frames as input data, which help eliminate useless body parts. We further model spatiotemporal relations based on the differential values of body parts. It is true that human actions are consistent in magnitude and frequency. In other words, there will be a trend in variations of position values. By describing trends of actions with the differential values of body parts and keeping trend information in the memory cell, the output of the R-LSTM unit could be more convinced and robust.

The structure of the proposed R-LSTM is represented in Figure 4A. Note that R-LSTM consists of the relation-aware part R and the typical LSTM part N, where R is constructed to describe spatial relations between body parts and represented in Figure 4B. Following the work of Shahroudy et al,[22] we divide the human body joints into five parts $P = \{p_k \mid k = 1, \ldots, 5\}$, where p_k is the set of corresponding body joints j_i. The formulations for the R-LSTM unit thus could be written as follows:

$$\begin{pmatrix} n_k^i \\ n_k^f \\ n_k^g \end{pmatrix} = \begin{pmatrix} \text{Sigm} \\ \text{Sigm} \\ \text{Tanh} \end{pmatrix} \left(W_k^n \begin{pmatrix} p_k^t \\ p_k^t - p_k^{t-1} \\ h_k^{t-1} \end{pmatrix} \right) \tag{2}$$

$$c_k^t = \left(\alpha r_k^f + (1-\alpha) n_k^f \right) \odot c_k^{t-1} + \alpha \left(r_k^i \odot r_k^g \right) + (1-\alpha) \left(n_k^i \odot n_k^g \right) \tag{3}$$

$$o = \text{Sigm} \left(W_o \cdot \left(p_1^t, \ldots, p_K^t, r_1^t, \ldots, r_K^t, h^{t-1} \right)^T \right) \tag{4}$$

$$h^t = o \odot \text{Tanh} \left(c_1^t, \ldots, c_K^t \right)^T, \tag{5}$$

where T refers to the transpose operation for the matrix, W_k^n and W_o represent the learned weight matrices, and α is a preset weight for the relation-aware part R. Essentially, Equation (2) represents that, in the typical LSTM part N, input gate n_k^i, forget gate n_k^f, and input modulation gate n_k^g corresponding to the kth body part are determined by the positions p_k^t, the difference in positions $p_k^t - p_k^{t-1}$ between time t and $t-1$, and former output state h_k^{t-1}. Equation (3) describes that the

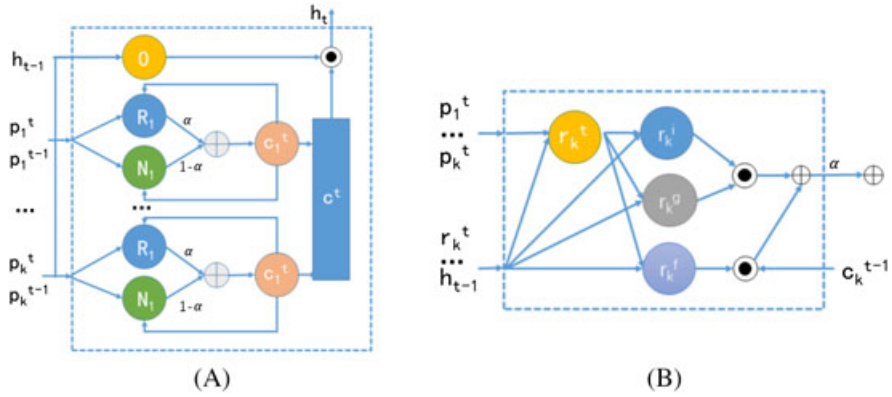

FIGURE 4 The structure of the relation-aware long short-term memory unit and its relation-aware part R are represented in (A) and (B), respectively. Note that N denotes the typical long short-term memory part [Color figure can be viewed at wileyonlinelibrary.com]

retained information of the internal memory cell c_k^t is a combination of the former memory after forgetting, information generated from the spatial relation of body parts, and information generated from input. Meanwhile, Equation (4) computes the output based on positions p_k^t, difference in positions of body parts r_k^t, and former output state h^{t-1}, which is determined by output o and internal memory cell state c_k in Equation (5).

The structure of the relation-aware part R is represented in Figure 4B, which can be formulated as follows:

$$r_k^t = \bigcup_{i=1}^{K} \tanh\left(W_k^i p_k^t - p_i^t\right), \text{where } i \neq k \tag{6}$$

$$\begin{pmatrix} r_k^i \\ r_k^f \\ r_k^g \end{pmatrix} = \begin{pmatrix} \text{Sigm} \\ \text{Sigm} \\ \text{Tanh} \end{pmatrix} \left(W_k^r \cdot \begin{pmatrix} r_k^t \\ r_k^t - r_k^{t-1} \\ h_k^{t-1} \end{pmatrix} \right), \tag{7}$$

where $\bigcup$ represents the concatenate operation and W_k^r is a learned weight matrix. We notice that Equation (6) utilizes the weighted difference between the kth body part and other body parts to form the spatiotemporal relation descriptor r_k^t. Meanwhile, r_k^t is adopted to construct the input gate r_k^i, forget gate r_k^f, and input modulation gate r_k^g of the relation-aware part in Equation (7). The constructed r_k^i, r_k^f, and r_k^g would affect the internal memory of R-LSTM, as illuminated in Equation (3). After constructing the R-LSTM network, we extract the corresponding feature in the softmax layer as R-LSTM feature F_l, which represents spatiotemporal relation between body parts.

6 | FUSION OF HETEROGENEOUS FEATURES BY AE NETWORK

In this section, we propose to fuse heterogeneous features, ie, the constructed WTP and R-LSTM features, to generate a more discriminative feature for action recognition. It is noted that we fuse heterogeneous features due to the fact that objects usually have heterogeneous representations and that researchers could learn their correlations at a "mid-level"[62] to help improve the robustness and correctness of recognition by fusing different representations of objects. Inspired by the work of Wu et al,[63] which fuses multimodal data, ie, RGB and depth, to learn a shared representation for gesture segmentation and recognition, we generate two different kinds of features from raw skeleton data, ie, WTP feature F_w and R-LSTM feature F_l, to fuse a distinctive feature for accurate action recognition.

Different from the work of Wu et al,[63] which uses a 3D convolutional neural network and stacked restricted Boltzmann machine/deep belief network to represent features before fusion, we adopt R-LSTM and WTP instead, and the architecture of the proposed fusing model is presented in Figure 5. The main reason for adopting an AE for fusion lies in the fact that an AE network is a natural and highly effective way to encode and decode information, especially for multimodal and heterogeneous features. During the process of encoding, the informative part of information can be merged for a higher distinctive representation, which has been proved effective by many methods and applications.[11,62] To speed up fuse operation, we argue that "prefused" weights could be directly used as initializations for the AE network, due to goal consistency, ie, assigning labels to human actions, between former steps and the fusing step. Specifically, we adopt a pretrained fully connected network accompanied with a small data set D, which not only assigns initial weights ω_w for WTP but also helps reduce dimensions of the WTP feature for a compact representation.

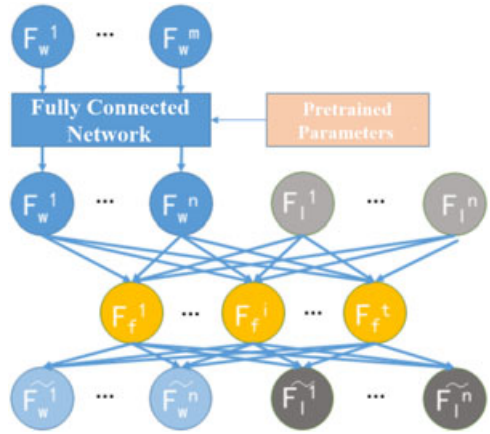

FIGURE 5 Architecture of the proposed fusion model, the basis of which being the autoencoder network. It is noted that we adopt pretrained parameters for the wavelet temporal pattern before fusing, which helps increase the convergence speed during training [Color figure can be viewed at wileyonlinelibrary.com]

In fact, the idea of adopting a fully connected network to settle initial parameters is similar to the spirit of a fully connected layer of LSTM, which successfully helps in transforming the initial weighting process into being one fully connected layer. We also take the same operation on the initial weights of R-LSTM ω_l, which is directly achieved from the previously trained R-LSTM feature F_l.

Afterward, the joint training with the AE network adjusts the parameters to handle the heterogeneity and produces a more reliable estimate from the heterogeneous data. The whole process of generating fusion feature F_s thus could be defined as

$$\left\{\tilde{F}_w(e_i), \omega_d\right\} = f_\tau(F_w(e_i); D) \tag{8}$$

$$F_s(e_i) = f_\mu\left(\omega_d, F_l(e_i), \omega_l, \tilde{F}_w(e_i)\right), \tag{9}$$

where functions $f_\tau()$ and $f_\mu()$ represent the logistic regression and AE network and $\tilde{F_w}$ refers to the WTP feature after dimensionality reduction. Note that we keep $\tilde{F_w}$ and F_l similar in dimension for equal representations. The training of the AE network ends when the validation error rate stops decreasing. During experiments, we find that our fusing model could end in less than 10 epochs, which proves the efficiency of our fusing model by adopting prefused weights. After fusing, we apply F_s in a logical regression model to get the label of action as $L = f_\tau(F_s(e_i))$.

7 | EXPERIMENTS

7.1 | Data sets

We evaluate our method on three public data sets, ie, UT-Kinect data set,[33] Florence 3D actions data set,[64] and NTU RGB+D data set,[22] where Table 1 offers the detailed descriptions of these three testing data sets.

UT-Kinect data set collects data via a stationary Kinect depth camera at a frame rate of 15 fps, containing RGB, Depth, and 3D skeleton data. UT-Kinect classifies samples into 10 kinds

TABLE 1 Detailed information on UTKinect, Florence 3D, and NTU RGB+D data sets

Name	Samples	Categories	Persons	Views	Description
UTKinect	199	10	10	1	RGB+Depth+3D Skeleton
Florence 3D	215	9	10	1	RGB+3D Skeleton
NTU RGB+D	56 880	60	40	80	RGB+Depth+3D Skeleton+Infrared Information

of daily-life actions, including "walking," "sitting down," "standing up," and so on. These actions are performed by 10 different persons with two trials of the same action. To sum up, a total of 199 action sequences are contained in this data set. Note that one of the original actions is invalid. The frame size in the UTKinect data set is different, varying from 5 to 120 frames. The UTKinect data set is challenging due to its wide intraclass differences and occlusion of body parts. For example, some of the "picking up things" actions are performed by either left- or right-hand persons, whereas others are done by using both hands. Generally speaking, there are two kinds of methods for validation in action recognition, ie, leave-one-out cross validation and 2-fold cross validation. We follow the idea in the works of Liu et al,[13] Xia et al,[33] and Hu et al[53] to utilize leave-one-out cross validation for the experiments.

Florence 3D data set collects data through a stationary Kinect as well, collecting nine common indoor action categories, such as "watching," "drinking water," "calling," and so on. Among these actions, nine actions are completed by 10 people, and each of the actions is performed repeatedly for 2 or 3 times, which sums up to a total of 215 actions. Compared with the UT-Kinect data set, the Florence 3D data set not only suffers from large intraclass differences but also is difficult in less inter variations between different classes. For example, "watching," "drinking water," and "calling" are similar from the perspective of the skeleton action sequence. Note that we use leave-one-out cross validation for the experiments, where we divided samples of the data set into 10 parts by action performer instead of randomly taking one part from 10 partitions.

NTU RGB+D data set is quite large in size compared with the former two data sets and collects four categories of data, including RGB, Depth, 3D skeleton, and infrared data. Each action sequence is captured by three stationary Kinect cameras, with cameras on both sides settled at an angle of 45 degrees to the middle one. Note that the size of the captured skeleton points in the NTU RGB+D data set is 25, which is larger than 15 skeleton points in the former two data sets. Over 40 people with ages from 10 to 35 years have completed 60 types of indoor actions, which sums up to a total of 56 880 action samples. Unlike UT-Kinect and Florence 3D, NTU RGB+D also designs one category of joint action performed by two persons. To deal with such cases, we directly splice the skeleton data of two persons as one skeleton sequence for experiments. We utilize the same verification method as in the work of Liu et al[13] for the cases of "cross subject," ie, half of the subjects for training and the other half for testing, and "cross view," ie, two viewpoints for training and the other one for testing, respectively.

7.2 | Experimental result analysis

Tables 2, 3, and 4 give the detailed statistics of our method and of other competing methods on NTU RGB+D, UT-Kinect, and Florence actions, respectively. It is noted that we apply the WTP feature and the R-LSTM feature for separate experiments to prove the effectiveness of the fusion of these two features for action recognition. Therefore, WTP and R-LSTM represent the detection results by only adopting the proposed WTP and R-LSTM features for classification in three Tables.

TABLE 2 Experimental results on the NTU RGB+D data set

Method	Cross Subject, %	Cross View, %
Proposed	**73.8**	**80.9**
WTP	70.1	77.5
R-LSTM	69.6	70.5
Du et al[40]	59.1	64.0
Liu et al[13]	69.2	77.7
Shahroudy et al[22]	62.9	70.3
Hu et al[65]	60.2	65.2

Abbreviations: R-LSTM, relation-aware long short-term memory; WTP, wavelet temporal pattern.

TABLE 3 Experimental results on the UT-Kinect data set

Method	Accuracy, %
Proposed	93.0
WTP	89.3
R-LSTM	90.4
Hu et al[53]	87.9
Liu et al[13]	**97.0**
Xia et al[33]	90.9

Abbreviations: R-LSTM, relation-aware long short-term memory; WTP, wavelet temporal pattern.

TABLE 4 Experimental results on the Florence actions data set

Method	Accuracy, %
Proposed	91.3
WTP	81.5
R-LSTM	88.3
Vemulapalli et al[10]	90.9
Anirudh et al[66]	89.7
Wang et al[67]	**91.6**

Abbreviations: R-LSTM, relation-aware long short-term memory; WTP, wavelet temporal pattern.

According to the fuse results from the three data sets, we conclude that fusion helps improve recognition accuracy greatly. We calculate that fusion increases the average accuracy from 79.6% by WTP and 79.7% by R-LSTM to 84.8% by the proposed method. This is intuitive since the robustness for detection is highly increased by adopting both temporal patterns and spatiotemporal relation features, other than using only one kind of feature. Moreover, the increase in accuracy proves the correctness and effectiveness of our fusion architecture.

We find that WTP and R-LSTM achieve an inconsistent performance when dealing with different data sets. For example, WTP achieves 77.5% on cross-view accuracy of the NTU RGB+D data set, which is much higher than 70.5% achieved by R-LSTM. Meanwhile, LSTM gets 88.3% on the Florence 3D actions data set, which is much higher than 81.5% achieved by WTP. We conclude that this is due to the different action categories contained in each data set. More detailed, the action categories in the Florence 3D actions data set are likely in shape of joint trajectories, such as "drink," "answer phone," and "check time." WTP cannot deal with the slight changes in actions since the main focus of WTP is to distinguish temporal patterns in a global manner, whereas R-LSTM keeps information of spatial relations between each frame, which helps distinguish slight variances. On the contrary, keeping information between frames makes it easy to confuse between locally plausible actions, which results in lower accuracy by R-LSTM compared with WTP.

Jointly learning WTP and R-LSTM leads to the consistent and high-accuracy performance achieved on the three data sets, which demonstrates the effectiveness and generality of the proposed method. More detailed, our method achieves the highest 73.8% and 80.9% on the challenging NTU RGB+D data set, the second highest 93.0% on the UT-Kinect data set, and the almost equally highest 91.3% on the Florence 3D actions data set. By incorporating temporal patterns and spatiotemporal relation, our method even outperforms several full LSTM methods in accuracy. For example, the accuracy on the NTU RGB+D data set by the proposed method is average 77.4% compared with average 73.5% achieved by Liu et al.[13] This proves the effectiveness of incorporating temporal patterns to improve recognition accuracy in a global manner. However, we find that the proposed method is low in accuracy for the Florence 3D actions and UT-Kinect data sets; LSTM needs quantity of training examples. However, these two data sets are small ones with only 200 and 215 action sequences compared to NTU RGB-D, which consists of 56 000 action sequences. Essentially, the failure cases on the Florence 3D actions and UT-Kinect data sets can be related to having not enough training samples; meanwhile, failure cases on the NTU RGB+D data set lie in its complicated interclass patterns with different persons, views, and other factors. Further improvements, such as the attention model or ResNet, can be made to increase the accuracy by enhancing the distinctive ability of extracted features, which is our plan for future work.

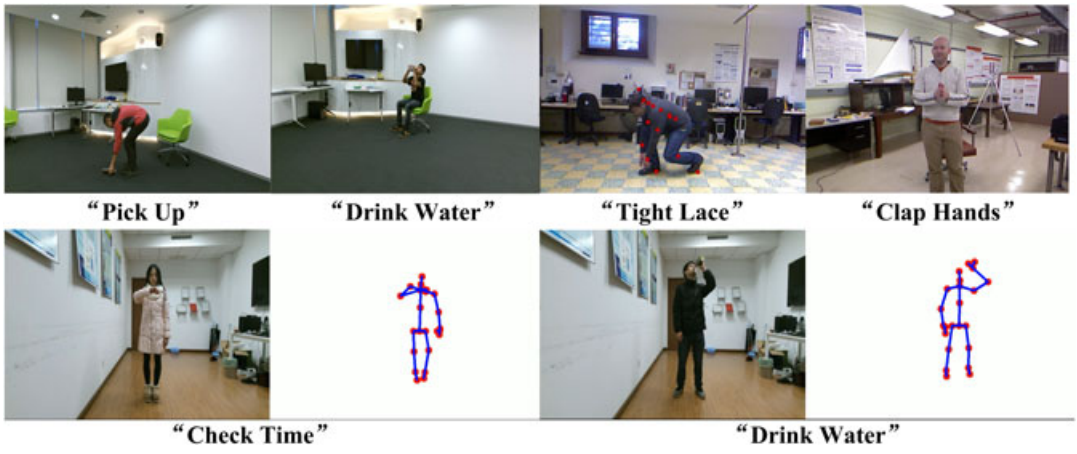

FIGURE 6 Action recognition examples of the proposed method on the NTU RGB+D, UT-Kinect, and Florence 3D actions and our captured action sequences. Note that action recognition results are given under double quotes [Color figure can be viewed at wileyonlinelibrary.com]

Several action recognition examples on the three data sets are shown in Figure 6, where the first and second rows show the recognition results on samples from the three data sets and our own captured action sequences, respectively. From these sample results, we can see that the proposed method could accurately recognize actions, even facing challenges of diversity and complexity in actors and the layout of background images. Moreover, the fusion of WTP and R-LSTM features could enhance the discrimination ability of the generated features, which leads to a better recognition result.

7.3 | Implementation details

The proposed method is implemented with Keras architecture and run on a laptop (2.6-GHz 4-core CPU, 16G RAM, Nvidia GTX 960M, and Windows 64-bit OS) for all the experiments. In order to retain more information on each body part, we repeat shoulder joints and hip joints. Hence, each action has more than eight joints. Our R-LSTM model includes two parts: R-LSTM layer and softmax layer. In the R-LSTM layer, the parameter α is assigned 0.3, the optimizer is RMSprop, and the learning rate is 0.01. We choose the optimum of α by experiments. In detail, we randomly choose 500 action sequences from our data sets, ie, NTU, Florence 3D, and UTK, to determine the optimal value. We plot a graph for recognition rate versus different α values. According to the experiments, the value for α is finally selected as 0.3.

8 | CONCLUSIONS

In this paper, we have proposed a robust 3D action recognition method by joint learning the temporal patterns and spatiotemporal relations of body joints. We first propose WTP to model temporal patterns in the time-frequency domain, which adaptively divides an action into subactions and extracts convinced representations in temporal patterns for subactions. The proposed R-LSTM is then proposed to model the strong dependency between body parts in the spatiotemporal domain. Regarding WTP and R-LSTM features as heterogeneous representations for actions, we finally fuse both features to define a robust and discriminative descriptor for action recognition. Experiments on three public data sets have shown the power of the proposed method in accurately and robustly recognizing actions performed by various actors. We believe that our proposed method could be utilized in many vision-based applications after optimization,[68] such as ill-health and computer-human interaction. In the future work, we aim to develop such software to further expand its usage and applicable scenarios.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under grant 2018YFC0407901; the National Natural Science Foundation of China under grants 61702160, 61872219, and 61702277; the Fundamental Research Funds for the Central Universities under grant 2016B14114; the Natural Science Foundation of Jiangsu Province under grant BK20170892; and the National Key Laboratory for Novel Software Technology, NJU, under grant K-FKT2017B05.

ORCID

Yirui Wu https://orcid.org/0000-0003-3022-3718

REFERENCES

1. Chang X, Yu Y-L, Yang Y, Xing EP. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(8):1617-1632.
2. Chang X, Yang Y, Xing EP, Yu Y-L. Complex event detection using semantic saliency and nearly-isotonic SVM. In: Proceedings of the 32nd International Conference on Machine Learning (ICML); 2015; Lille, France.
3. Chang X, Yu Y-L, Yang Y, Xing EP. They are not equally reliable: semantic event search using differentiated concept classifiers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV.
4. Wang H, Chang X, Shi L, Yang Y, Shen Y-D. Uncertainty sampling for action recognition via maximizing expected average precision. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018; Stockholm, Sweden.
5. Yang X, Tian Y. Effective 3D action recognition using eigenjoints. *J Vis Commun Image Represent*. 2014;25(1):2-11.
6. Jing L, Yang X, Tian Y. Video you only look once: overall temporal convolutions for action recognition. *J Vis Commun Image Represent*. 2018;52:58-65.
7. Chang X, Ma Z, Lin M, Yang Y, Hauptmann AG. Feature interaction augmented sparse learning for fast kinect motion detection. *IEEE Trans Image Process*. 2017;26(8):3911-3920.
8. Dinerstein J, Ventura D, Egbert PK. Fast and robust incremental action prediction for interactive agents. *Computational Intelligence*. 2005;21(1):90-110.
9. Shao L, Zhen X, Tao D, Li X. Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Trans Cybern*. 2014;44(6):817-827.
10. Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014; Columbus, OH.
11. Shahroudy A, Ng T-T, Gong Y, Wang G. Deep multimodal feature analysis for action recognition in RGB+ D videos. *IEEE Trans Pattern Anal Mach Intell*. 2018;40(5):1045-1058.
12. Zhu W, Lan C, Xing J, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016; Phoenix, AZ.
13. Liu J, Shahroudy A, Xu D, Wang G. Spatio-temporal LSTM with trust gates for 3D human action recognition. In: *Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*. Cham, Switzerland: Springer International Publishing; 2016.
14. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2005; San Diego, CA.
15. Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia (MM); 2007; Augsburg, Germany.
16. Willems G, Tuytelaars T, Van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector. In: *Computer Vision - ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*. Berlin, Germany: Springer-Verlag Berlin Heidelberg; 2008.
17. Wang H, Kläser A, Schmid C, Liu CL. Dense trajectories and motion boundary descriptors for action recognition. *Int J Comput Vis*. 2013;103(1):60-79.
18. Abedinia O, Amjady N, Ghadimi N. Solar energy forecasting based on hybrid neural network and improved metaheuristic algorithm. *Computational Intelligence*. 2018;34(1):241-260.
19. Abbaszadeh P, Alipour A, Asadi S. Development of a coupled wavelet transform and evolutionary Levenberg-Marquardt neural networks for hydrological process modeling. *Computational Intelligence*. 2018;34(1):175-199.
20. Wei L, Wu Y, Wang W, Lu T. A novel 3D human action recognition framework for video content analysis. In: *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I*. Cham, Switzerland: Springer International Publishing; 2018.

21. Shahroudy A, Ng T-T, Yang Q, Wang G. Multimodal multipart learning for action recognition in depth videos. *IEEE Trans Pattern Anal Mach Intell*. 2016;38(10):2123-2129.
22. Shahroudy A, Liu J, Ng T-T, Wang G. NTU RGB+ D: a large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV.
23. Hu J-F, Zheng W-S, Pan J, Lai J, Zhang J. Deep bilinear learning for RGB-D action recognition. In: Proceedings of the 15th European Conference on Computer Vision (ECCV); 2018; Munich, Germany.
24. Ho ESL, Chan JCP, Chan DCK, Shum HPH, Cheung Y, Yuen PC. Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. *Comput Vis Image Underst*. 2016;148:97-110.
25. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y. Robust 3D action recognition with random occupancy patterns. In: *Computer Vision - ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*. Berlin, Germany: Springer-Verlag Berlin Heidelberg; 2012.
26. Lu C, Jia J, Tang C-K. Range-sample depth feature for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014; Columbus, OH.
27. Oreifej O, Liu Z. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2013; Portland, OR.
28. Evangelidis G, Singh G, Horaud R. Skeletal quads: human action recognition using joint quadruples. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR); 2014; Stockholm, Sweden.
29. Zou B, Chen S, Shi C, Providence UM. Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. *Pattern Recognition*. 2009;42(7):1559-1571.
30. Wang T, Xu R. Some integral inequalities in two independent variables on time scales. *J Math Inequalities*. 2012;6(1):107-118.
31. Zong Z, Hu F. Lp solutions of infinite time interval backward doubly stochastic differential equations. *Filomat*. 2017;31(7):1857-1868.
32. Liu H. A class of retarded Volterra-Fredholm type integral inequalities on time scales and their applications. *J Inequalities Appl*. 2017;2017(1):293.
33. Xia L, Chen C-C, Aggarwal JK. View invariant human action recognition using histograms of 3D joints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2012; Providence, RI.
34. Ohn-Bar E, Trivedi M. Joint angles similarities and HOG2 for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2013; Portland, OR.
35. Liang C, Chen E, Qi L, Guan L. Action recognition using multi-layer depth motion maps and sparse dictionary learning. Paper presented at: 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP); 2015; Xiamen, China.
36. Fernando B, Gavves E, Oramas J, Ghodrati A, Tuytelaars T. Rank pooling for action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(4):773-787.
37. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Computation*. 2006;18(7):1527-1554.
38. Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*. 2017;68:346-362.
39. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F. A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI.
40. Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA.
41. Veeriah V, Zhuang N, Qi G. Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015; Santiago, Chile.
42. Wang H, Wang L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. CoRR. 2017. https://arxiv.org/abs/1704.02581

43. Chen L, Zhang H, Xiao J, et al. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. 2016. arXiv preprint arXiv:1611.05594.
44. Woo S, Park J, Lee J-Y, So Kweon I. CBAM: convolutional block attention module. In: Proceedings of the 15th European Conference on Computer Vision; 2018; Munich, Germany.
45. Song S, Lan C, Xing J, Zeng W, Liu J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence; 2017; San Francisco, CA.
46. Liu J, Wang G, Hu P, Duan L-Y, Kot AC. Global context-aware attention LSTM networks for 3D action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI.
47. Xu X, Fu S, Qi L, et al. An IoT-oriented data placement method with privacy preservation in cloud environment. *J Netw Comput Appl*. 2018;124:148-157.
48. Qi L, Zhang X, Dou W, Ni Q. A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data. *IEEE J Sel Areas Commun*. 2017;35(11):2616-2624.
49. Xu X, Huang R, Dou R, et al. Energy-efficient cloudlet management for privacy preservation in wireless metropolitan area networks. *Secur Commun Netw*. 2018;2018:8180451:1-8180451:13.
50. Gong W, Qi L, Xu Y. Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. *Wirel Commun Mob Comput*. 2018;2018.
51. Xu X, Zhao X, Ruan F, et al. Data placement for privacy-aware applications over big data in hybrid clouds. *Secur Commun Netw*. 2017;2017:2376484:1-2376484:15.
52. Zhu Y, Chen W, Guo G. Fusing spatiotemporal features and joints for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2013; Portland, OR.
53. Hu J-F, Zheng W-S, Lai J, Zhang J. Jointly learning heterogeneous features for RGB-D activity recognition. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(11):2186-2200.
54. Liu L, Shao L. Learning discriminative representations from RGB-D video data. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI); 2013; Beijing, China.
55. Ni B, Wang G, Moulin P. RGBD-HuDaAct: a color-depth video database for human daily activity recognition. In: *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. London, UK: Springer; 2013:193-208.
56. Song Y, Liu S, Tang J. Describing trajectory of surface patch for human action recognition on RGB and depth videos. *IEEE Signal Process Lett*. 2015;22(4):426-429.
57. Kong Y, Fu Y. Bilinear heterogeneous information machine for RGB-D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA.
58. Keogh E, Ratanamahatana CA. Exact indexing of dynamic time warping. *Knowl Inf Syst*. 2005;7(3):358-386.
59. Wang J, Liu Z, Wu Y, Yuan J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2014;36(5):914-927.
60. Jiang J, Liu L. Existence of solutions for a sequential fractional differential system with coupled boundary conditions. *Bound Value Probl*. 2016;2016(1):159.
61. Meng F, Feng Q. A new fractional subequation method and its applications for space-time fractional partial differential equations. *J Appl Math*. 2013;2013.
62. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML); 2011; Bellevue, WA.
63. Wu D, Pigou L, Kindermans P-J, et al. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans Pattern Anal Mach Intell*. 2016;38(8):1583-1597.
64. Seidenari L, Varano V, Berretti S, Bimbo A, Pala P. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2013; Portland, OR.
65. Hu J-F, Zheng W-S, Lai J, Zhang J. Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA.
66. Anirudh R, Turaga P, Su J, Srivastava A. Elastic functional coding of human actions: from vector-fields to latent variables. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015; Boston, MA.

67. Wang P, Yuan C, Hu W, Li B, Zhang Y. Graph based skeleton motion representation and similarity measurement for action recognition. In: *Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. Cham, Switzerland: Springer International Publishing; 2016.
68. Yuan Y, Xu H, Wang B, Yao X. A new dominance relation-based evolutionary algorithm for many-objective optimization. *IEEE Trans Evol Comput*. 2016;20(1):16-37.

How to cite this article: Wu Y, Wei L, Duan Y. Deep spatiotemporal LSTM network with temporal pattern feature for 3D human action recognition. *Computational Intelligence*. 2019;1–20. https://doi.org/10.1111/coin.12207