Mathematical Biosciences
and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(6): 11313–11327.
DOI: 10.3934/mbe.2023501
Received: 16 March 2023
Revised: 15 April 2023
Accepted: 19 April 2023
Published: 26 April 2023

Research article

Efficient motion capture data recovery via relationship-aggregated graph network and temporal pattern reasoning

Chuanqin Zheng[1], **Qingshuang Zhuang**[1,2] **and Shu-Juan Peng**[2,*]

[1] Information Center, Xiamen Medical College, Xiamen, China
[2] Department of Artificial Intelligence, Huaqiao University, Xiamen, China

* **Correspondence:** Email: pshujuan@hqu.edu.cn; Tel: +86-592-6162556.

Abstract: Human motion capture (mocap) data is of crucial importance to the realistic character animation, and the missing optical marker problem caused by marker falling off or occlusions often limit its performance in real-world applications. Although great progress has been made in mocap data recovery, it is still a challenging task primarily due to the articulated complexity and long-term dependencies in movements. To tackle these concerns, this paper proposes an efficient mocap data recovery approach by using Relationship-aggregated Graph Network and Temporal Pattern Reasoning (RGN-TPR). The RGN is comprised of two tailored graph encoders, local graph encoder (LGE) and global graph encoder (GGE). By dividing the human skeletal structure into several parts, LGE encodes the high-level semantic node features and their semantic relationships in each local part, while the GGE aggregates the structural relationships between different parts for whole skeletal data representation. Further, TPR utilizes self-attention mechanism to exploit the intra-frame interactions, and employs temporal transformer to capture long-term dependencies, whereby the discriminative spatio-temporal features can be reasonably obtained for efficient motion recovery. Extensive experiments tested on public datasets qualitatively and quantitatively verify the superiorities of the proposed learning framework for mocap data recovery, and show its improved performance with the state-of-the-arts.

Keywords: mocap data recovery; relationship-aggregated graph network; temporal pattern reasoning; self-attention mechanism

1. Introduction

Mocap (mocap) technology aims to capture highly precise recordings of the real movements, which is popularized in a variety of purposes including computer games, augmented reality, movie production, human-computer interactions and so forth [1]. Often, these systems utilize a large number of attached optical markers to record the human motion track [2]. Nevertheless, even with the highly professional capturing equipments and the most sophisticated software available, the current mocap systems still suffer from the incomplete motion recordings due to the sensor noise, marker falling off or occlusion problem. Prior to use, some users select to simply discard the missing recordings. However, the simple removal of these missing information often fail to retain a continuous data flow and may lose lots of valuable information significantly.

In general, human mocap data is acquired by the highly articulated movements that collected from the fixed markers [3], and it is reasonable to utilize the kinematic constraints within the human skeletal data to handle the missing marker problem. For instance, the surrounding markers that share the kinematic relationships with the missing markers is capable of assisting the missing entry estimation. Along this line, some filtering techniques and motion matrix analysis approaches have been proposed to restore the missing joints. Nevertheless, it is still a challenging task to automatically recovery the missing joints, and the main reasons are three-fold: 1) Non-linear property: the articulated human motions are always non-linearly correlated, which make it difficult to learn the underling data structure for motion recovery; 2) Randomicity: the distribution of missing joints is often unknown and arbitrary, which make it hard to adaptively recover the missing joints. 3) Long-term dependency: the human motions always contain various kinds of different actions, and the long-term temporal dependencies with a large part of missing markers may not be easily captured for precise motion recovery.

The spatial dependency of skeleton joints in the current frame and the temporal dependency of the same joint among the neighboring frames are of crucial importance to the mocap data analysis. In recent years, some research works have attempted different deep neural networks to simulate the spatio-temporal correlation within the human motions, and propose various models to recover the missing joints, e.g., convolutional autoencoder [4] and bi-directional long short-time memory network (BLSTM) [5]. Although these approaches have achieved significant performances, their recovery accuracies may degrade rapidly over a long period of motion sequence and still face two challengings: 1) The skeleton joints within the human motions are articulated with each other, and it is difficult to learn the spatio-temporal features in a reliable way. 2) The current deep models often fail to capture the long-term temporal dependency due to the accumulated training errors.

Until recently, the graph models are effectively to represent the objects and their relationships interpretably, and have promptly become a powerful tool in high-level representation understanding tasks [6,7]. Inspired by the great success of graph models that can flexibly learn the high-level semantic information, this paper proposes an efficient mocap data recovery approach by using Relationship-aggregated Graph Network and Temporal Pattern Reasoning (RGN-TPR). In summary, the proposed framework provides the following three contributions:

- An efficient relationship-aggregated graph model is proposed to aggregate both of the local node-level relationship and global body-level relationship within the human body, which can be well utilized to discriminatively model the human skeletal data.
- A temporal pattern reasoning module is efficiently addressed to exploit the intra-frame interactions and temporal correlation within the human motion, whereby the temporal dynamics and long-term dependency can be well obtained for efficient motion recovery.
- Extensive experiments verify the superiorities of the proposed framework under various motion recovery tasks, and show its improved performance over the state-of-the-arts.

The remaining part of this paper is organized as follows: Section 2 surveys the existing mocap data recovery works, and Section 3 introduces the proposed framework in detail. The experiments and comparisons with the state-of-the-arts are stated in Section 4. Finally, we draw a conclusion in Section 5.

2. Related work

Human mocap data primarily records the 3D position and orientation information about the moving body, and the missing marker problem often degrades the motion quality. In the past, various mocap data recovery methods have been developed, which can be broadly divided into two branches: data-structure based methods and data-driven based methods.

2.1. Data-structure based methods

The data-structure based methods mainly employs the neighboring available markers or statistical structure property to recover the missing joints, which have the advantages of low computational cost and fast implementation. Intuitively, motion filtering is able to complete the missing values in the data sequence. Along this line, Ristidou et al. [8] utilize the relevant information from neighboring markers and select Kalman filter to supplement continuous flow related to rigid body motions. Similarly, Wu et al. [9] propose a piecewise linear Kalman filter to predict the location of missing markers, while Burke et al. [10] combine the Kalman filter and motion smoothing operation to recover the missing markers in a low dimensional Kalman smoothing space. Note that, these methods generally need to define the inherent kinematic constraints within the available joint information in advance, which often tend to produce unreasonable results for significantly corrupted motion sequences. Besides, these approaches are often performed with the assistance of a significant amount of human intervention, which inevitably require the manual tuning of the filter parameters to handle different motions.

Alternatively, the missing marker problem has been traditionally formulated as a matrix completion task, which mainly explore the linear or non-linear properties of motion matrix to restore the missing markers. For instance, Lai et al. [11] utilize low-rank prior to recover the damaged human motion matrix, in which the singular value threshold operation is selected to solve the rank minimization problem. Tan et al. [12] first divide the mocap data into a group of trajectory-based segments, and then perform local matrix completion to achieve missing entry estimation. Feng et al. [13] first consider the low-rank structure and the temporal properties of the motion data, and then utilize robust matrix completion problem to refine the recovered motion sequence. In addition, Peng et al. [14] decompose the underling human skeleton data into several blocks, and utilize the adaptive nonnegative matrix factorization to achieve incomplete human mocap data recovery. These approaches are able to well restore the simple motion with limited missing entries, but which may be unsuitable to tackle the complex motions with a large portion of missing joints for an extended period of time.

2.2. Data-driven based methods

Data-driven based approaches primarily learn the motion model from the available dataset to reconstruct the missing entries. Along this line, Herda et al. [15] exploit a sophisticated human model to learn a precise representation of the skeleton data, whereby the 3D location of markers can be well predicted. Li et al. [16] investigate some hidden variables from the observed mocap data and learn the latent variables to estimate the missing values. Recently, Xiao et al. [17] design a sparse representation of the incomplete observations and further utilize such sparse representation to predict the missing markers. These methods have been proved to be effective in recovering the missing entries, but which may change the topological structure within the raw mocap data such that some restored frames may exist unnatural movements.

With the rapid development of deep learning, there is a potential feasibility in applying deep learning techniques for synthesizing the complete motion data. For instance, Holden et al. [4] utilize the convolutional autoencoder to extract human motion manifold structure for motion synthesis. By considering the dynamic characteristics of mocap data, Cui et al. [5] embed the attention mechanism of bidirectional long-short term memory network (BLSTM) structure in the encoding and decoding stage, which can adaptively extract the relevant information at each time step for human motion recovery. Ji et al. [18] address a least square filtering to optimize the long-short term memory network and utilize the attention mechanism to model the temporal property of human mocap data. Although these deep models take into account the spatial-temporal characteristics of motion sequences, they still suffer from performance degradation with a significant portion of missing markers and the increase of sequence length. Therefore, it is necessary to develop a more robust deep model for mocap data recovery.

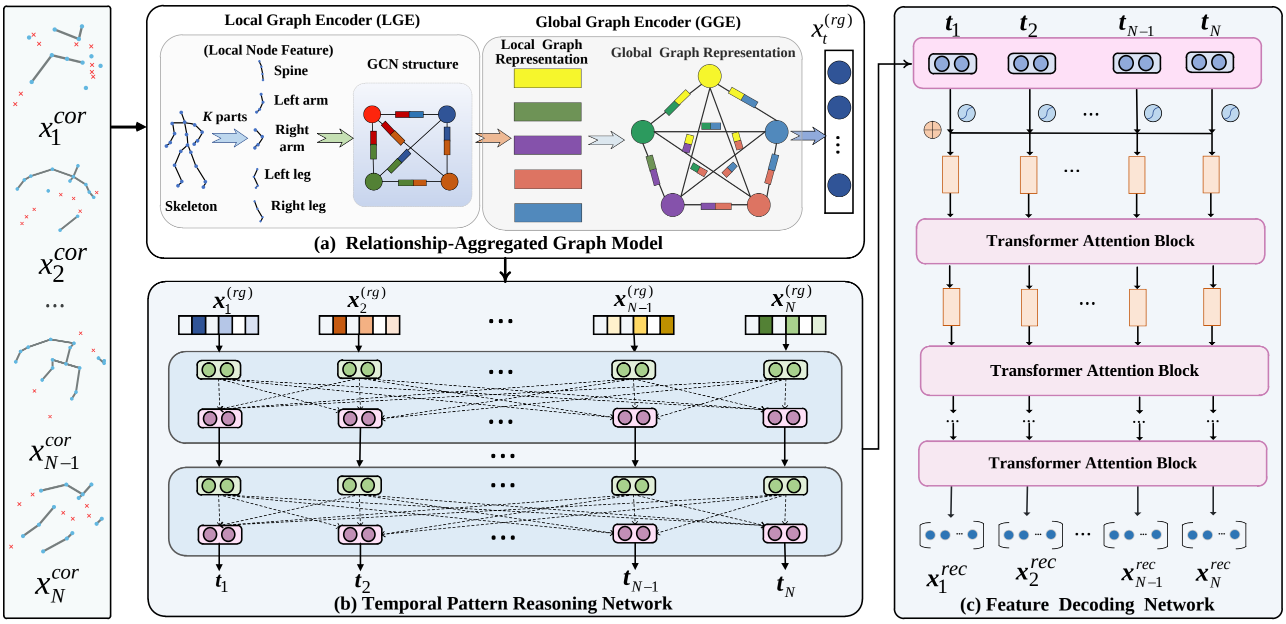


Figure 1. The pipeline of the proposed motion recovery framework.

3. Methodology

The human skeleton data is generally recorded by a group of articulated joints, and the articulated complexity of human mocap data often make it difficult for motion recovery. To address this issue, we present an efficient mocap data recovery approach by relationship-aggregated graph network and temporal pattern reasoning, and the architecture of the proposed framework is shown in Figure 1. The proposed framework is comprised of two tailored graph encoders, local graph encoder (LGE) and global graph encoder (GGE). By dividing the human skeletal structure into several parts, LGE module encodes the high-level semantic node features and their semantic relationships in each local part, while the GGE module aggregates the structural relationships between different parts for whole skeletal data representation. Further, the temporal pattern reasoning module utilizes self-attention mechanism to exploit the intra-frame interactions, and employs temporal transformer to capture long-term dependencies, whereby the discriminative spatio-temporal features can be reasonably obtained for efficient motion recovery. This section first clarifies the notations and formal definitions of motion recovery. Then, the proposed network architecture and its learning modules are introduced in tandem. Finally, the recovery of missing mocap data and its optimization process are explicitly provided.

3.1. Problem formulation and notation

Mocap data mainly consists of a group of motion frames, which are connected by highly articulated joints [19]. In general, each frame records the 3D position of every joint, and mocap data can be formulated as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{N\times 3J}$, where $\mathbf{x}_t \in \mathbb{R}^{1\times 3J}$ represent the t-th frame, J and N are respectively the joint number in the human skeleton model and the total frame number in the motion sequence. Without loss of generality, let $\mathbf{X}^{cor} = [\mathbf{x}_1^{cor}, \mathbf{x}_2^{cor}, \cdots, \mathbf{x}_N^{cor}] \in \mathbb{R}^{N\times 3J}$ be a corrupted motion sequence with missing joint values, we utilize a binary mask matrix $\mathbf{M} \in \mathbb{R}^{N\times 3J}$ to indicate the missing positions ($\mathbf{M}_{ij} = 0$) and normal positions ($\mathbf{M}_{ij} = 1$), and it is easily to obtain the data relationship $\mathbf{X}^{cor} = \mathbf{X} \odot \mathbf{M}$, where the symbol $\odot$ is element-wise product. Accordingly, the motion recovery problem can be transformed into optimizing the functions g and f to minimize the difference between the recovered motion $f(g(\mathbf{X}^{cor}))$ and the complete motion sequence $\mathbf{X}$:

$$\min_{f,g} ||\mathbf{X} - f(g(\mathbf{X}^{cor}))|| \tag{3.1}$$

where encoder $g(\mathbf{X}^{cor})$ is designed to map the observation $\mathbf{X}^{cor}$ into a high-level spatio-temporal representation, and the decoder $f(g(\mathbf{X}^{cor}))$ transforms the output back into the input manifold to reconstruct the original data.

3.2. Relationship-aggregated graph model

The skeleton data is always represented as a sequence of feature vectors, and each vector characterizes the 3D coordinates that are relevant to the corresponding human joint. Besides, the bone information, which represents the directions and lengths of bones, has also been proved to be valuable for skeleton-based motion analysis. In this section, a relationship-aggregated graph model, consisting of two tailored graph encoders, local graph encoder (LGE) and global graph encoder (GGE), are respectively proposed to encode the node-level relationship and body-level relationship for skeleton data representation.

3.2.1. Local graph encoder

Graphs are particularly effective in modeling the complex structured data, and graph convolutional network (GCN) has achieved remarkable performance mainly due to its efficient representation in modeling the dependencies in skeletal data. Nevertheless, some joints located in different parts of the body may not have the direct physical dependencies between each other, and the global topology of the graph model may not well characterize the complex mocap data due to its inherent articulated complexity. To tackle these problems, we decompose the underling human skeleton model into K parts (e.g., $K = 5$ referred to work [14]), and the partition groups are denoted as: $\mathcal{P}$={spine, left hand, right hand, left leg, right leg}. Let $\bar{\mathbf{x}}_t^p$ denote the feature representation of the nodes contained in the p-th partition, we construct a graph $\mathcal{G}^p(\mathbf{V}^p, \mathbf{E}^p)$ to characterize the p-th partition, where $\mathbf{V}^p \in \bar{\mathbf{x}}_t^p$ is a set of joint feature vector, $\mathbf{E}^p$ denotes the relationship among two connected joints and weighted by adjacent matrix $\mathbf{A}^{(p)} \in \mathbb{R}^{n_p \times n_p}$. Note that, the weight of the edge represents the relationship degreebetween two joints, and the weighted edge $\mathbf{e}_{ij}^p$ between the joint $\mathbf{v}_i^p$ and joint $\mathbf{v}_j^p$ is calculated as follows:

$$\mathbf{e}_{ij}^p = \mathrm{ReLU}((\mathbf{v}_i^p \mathbf{W}_1^p + \mathbf{b}_1^p)(\mathbf{v}_j^p \mathbf{W}_2^p + \mathbf{b}_2^p)^{\mathrm{T}}) \tag{3.2}$$

where $\{\mathbf{W}_1^p, \mathbf{b}_1^p, \mathbf{W}_2^p, \mathbf{b}_2^p\}$ are the trainable parameters. Accordingly, the weighted adjacent matrix is defined as:

$$\mathbf{A}_{(i,j)}^p = \begin{cases} \mathbf{e}_{ij}^p, & (i,j) \in \mathbf{E}^p \\ 0, & otherwise \end{cases}, \tag{3.3}$$

Accordingly, the local graph $\mathcal{G}_p$ can be well constructed for p-th partition, which can well model the joint relationships in each part. Meanwhile, each joint aggregates significant information that connected to each other. Therefore, these graph nodes are aggregated with other connected joints by:

$$\mathbf{v}_i^p = \mathrm{ReLU}(\sum_{j=1}^{n_p} \mathbf{A}_{(i,j)}^p \times (\mathbf{v}_j^p \mathbf{W}_3^p + \mathbf{b}_3^p) + \mathbf{v}_i^p) \tag{3.4}$$

where $\mathbf{v}_i^p$ is the i-th node in $\mathcal{G}_p$, and $\{\mathbf{W}_3^p, \mathbf{b}_3^p\}$ are the trainable convolution parameters. Further, these updated node features $\{\mathbf{v}_1^p, \mathbf{v}_2^p \ldots \mathbf{v}_{n_p}^p\}$ in $\mathcal{G}_p$ are further normalized with l_2 norm: $\mathbf{V}^p = \left\| \{\mathbf{v}_1^p, \mathbf{v}_2^p \ldots \mathbf{v}_{n_p}^p\} \right\|_2$. For the i-th frame, we fuse these relationship-aggregated node features into a single feature vector to characterize each body part:

$$\mathcal{V}_t^p = \mathrm{flatten}(\mathbf{V}^p) \tag{3.5}$$

3.2.2. Global graph encoder

The skeleton joints located in different body parts often coordinate with each other to form different poses. For example, walking requires not only the legs to walk, but also involves the swinging of the arms to coordinate the balance of the body. To model the correlation of different body parts, we proposes a global graph encoder to aggregate the relationships between different body parts. Since the human skeleton model is divided into K parts, we construct an undirected graph $\mathcal{G}^g = (\mathbf{V}^g, \mathbf{E}^g)$ to aggregate the correlation between these local parts, where $\mathbf{V}^g \in \mathcal{V}_t$ is the set of feature vectors derived from the local body parts, $\mathbf{E}^g$ denotes the relationship between two parts and weighted by adjacent matrix $\mathbf{A}^g \in \mathbb{R}^{K \times K}$. The weighted edge $\mathbf{e}_{ij}^g$ between the i-th part $\mathbf{v}_i^g$ and the j-th part $\mathbf{v}_j^g$ is computed by:

$$\mathbf{e}_{ij}^g = \mathrm{ReLU}((\mathbf{v}_i^g \mathbf{W}_1^g + \mathbf{b}_1^g)(\mathbf{v}_j^g \mathbf{W}_2^g + \mathbf{b}_2^g)^{\mathrm{T}}) \tag{3.6}$$

where $\{\mathbf{W}_1^g, \mathbf{b}_1^g, \mathbf{W}_2^g, \mathbf{b}_2^g\}$ are the trainable parameters to calculate the weights of graph edges. Accordingly, the weighted adjacent matrix can be defined as:

$$\mathbf{A}_{(i,j)}^g = \begin{cases} \mathbf{e}_{ij}^g, & i \neq j \\ 0, & otherwise \end{cases}, \tag{3.7}$$

The global graph is capable of modeling the semantic relationship between different body parts, and each graph node can aggregate mutually relevant important information by using weighted edges:

$$\mathbf{v}_i^g = \mathrm{ReLU}(\sum_{j=1}^{K} \mathbf{A}_{(i,j)}^g \times (\mathbf{v}_j^g \mathbf{W}_3^g + \mathbf{b}_3^g) + \mathbf{v}_i^g) \tag{3.8}$$

where $\{\mathbf{W}_3^g, \mathbf{b}_3^g\}$ are the trainable convolution parameters. Further, the relationship-aggregated body features are further normalized with l_2 norm: $\mathbf{V}^g = \left\|\mathbf{v}_1^g, \mathbf{v}_2^g \ldots \mathbf{v}_k^g\right\|_2$. Finally, we concatenate these relationship-aggregated body features to characterize the spatial structure of each frame:

$$\mathbf{x}_t^{(rg)} = f_s(concat(\mathbf{v}_1^g, \mathbf{v}_2^g \ldots \mathbf{v}_k^g)) \tag{3.9}$$

where $f_s(\cdot)$ is a linear layer. Accordingly, the motion sequences are encoded by a sequence of relationship-aggregated feature vectors $\mathbf{X}^{(rg)} = [\mathbf{x}_1^{(rg)}, \ldots, \mathbf{x}_N^{(rg)}]$.

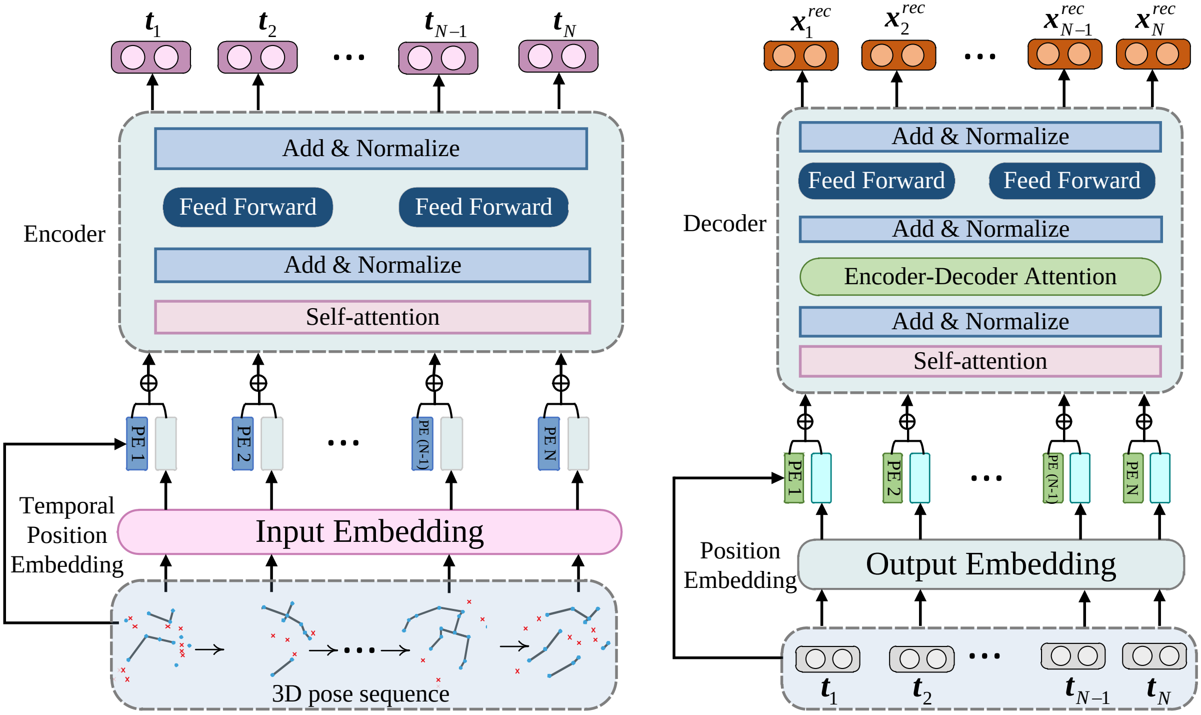


Figure 2. The architecture of the encoding and decoding networks. Left: Temporal pattern reasoning network for modeling the dynamic characteristics. Right: Feature decoding network for motion recovery.

3.3. Temporal pattern reasoning network

The temporal modeling of dynamic skeletons is of crucial importance to the discriminative motion analysis. On the one hand, the existing recurrent neural network (RNN) and long short-time memory network (LSTM) often fail to capture the long-term dependencies between the skeletons, mainly due to the uncertainty and diversity of human motion. Under such circumstances, the long-term temporal dependencies with a significant portion of missing markers may not be easily captured for precise motion recovery. On the other hand, not all frames contribute equally to the informative motion modeling, and it is necessary to focus on the informative frames in the sequence, and ignore the effects of the irrelevant frames. To alleviate these concerns, we propose a temporal pattern reasoning network (TPRN) to model the dependencies and dynamic information in the motions. As shown in Figure 2(a), a scaled multi-head self-attention layer is utilized to the updated node feature matrix $\mathbf{X}^{(rg)}$, and each head $h \in \{1, \ldots, \mathbf{h}\}$ is defined as:

$$\text{head}_h = \text{softmax}\left(\frac{\mathbf{X}_{Q_h}^{(rg)}(\mathbf{X}_{K_h}^{(rg)})^{\mathrm{T}}}{d}\right)\mathbf{X}_{V_h}^{(rg)} \tag{3.10}$$

where $\mathbf{X}_{Q_h}^{(rg)}$, $\mathbf{X}_{K_h}^{(rg)}$ and $\mathbf{X}_{V_h}^{(rg)}$ are head projections of feature matrix $\mathbf{X}^{(rg)}$ and d is a normalization factor. Note that, the softmax-function is often referred to as the attention weight matrix [20]. Finally, the updated feature matrix via multi-head self-attention mechanism is obtained by:

$$\text{MSA}(\mathbf{X}^{(rg)}) = (\text{head}_1 \parallel \cdots \parallel \text{head}_{\mathbf{h}})\,\mathbf{W}_o + \mathbf{b}_o \tag{3.11}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are trainable parameters for the outputs. Note that, the temploral self-attention mechanism relates different positions of input sequence, which can provide a simple and powerful reasoning mechanism to reason the hidden links between the vector entities. Let $\mathbf{Z}_1 = \text{MSA}(\mathbf{X}^{(rg)})$, the encoder is able to stack multiple instances of the same architecture, and the temporal transformer encoder with L layers is denoted as follows:

$$\mathbf{Z'}_l = \text{MSA}(\ell(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1}, l = 2, \ldots, L \tag{3.12}$$

$$\mathbf{Z}_l = \mathcal{M}(\ell(\mathbf{Z'}_l)) + \mathbf{Z'}_l, l = 2, \ldots, L \tag{3.13}$$

$$\mathbf{T} = \ell(\mathbf{Z}_L) \tag{3.14}$$

where $\ell(\cdot)$ denotes the layer normalization operator, and $\mathcal{M}(\cdot)$ is multilayer perceptron operator. Accordingly, the discriminative spatio-temporal features are denoted as: $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\} \in \mathbb{R}^{N\times d}$.

3.4. Feature decoding network

After all the frames are encoded, we can obtain a representation for the corrupted motion. As shown in Figure 2, a feature decoding network (FDN) is further designed to map the feature vector back into a recovered human motion. Similar to the encoding network, we also utilize multi-head self-attention mechanism to mine the most informative frames. Given the spatio-temporal features $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$, the FDN with L layers can be represented as follows:

$$\mathcal{T'}_l = \text{MSA}_d(\ell(\mathcal{T}_{l-1})) + \mathcal{T}_{l-1}, l = 2, \ldots, L \tag{3.15}$$

$$\mathcal{T}_l = \mathcal{M}_d(\ell(\mathcal{T'}_l)) + \mathcal{T'}_l, l = 2, \ldots, L \tag{3.16}$$

$$\mathbf{X}^{rec} = \ell(\mathcal{T}_L) \tag{3.17}$$

where $\text{MSA}_d(\cdot)$ and $\mathcal{M}_d(\cdot)$ are respectively the multi-head self-attention operation and multilayer perceptron operation in decoder network, which share the same network structure with the encoding process. Accordingly, the reconstructed motion sequence with the same size as the input motion can be obtained, i.e., $\mathbf{X}^{rec} = \{\mathbf{x}_1^{rec}, \mathbf{x}_2^{rec}, \ldots, \mathbf{x}_N^{rec}\}$. Finally, the recovered motion can be derived from the following formula:

$$\mathbf{X}^* = ||\mathbf{M} \odot \mathbf{X}^{cor} + (1 - \mathbf{M}) \odot \mathbf{X}^{rec}|| \tag{3.18}$$

where $\mathbf{X}^*$ is the weighted sum of $\mathbf{X}^{cor}$ and $\mathbf{X}^{rec}$. That is, only the missing joint is restored and the other parts are equal to the input.

3.5. Optimization

The encoding and decoding process enhances the robustness of the motion recovery results. It should be noted that the refined results should semantically match the input motion to maintain the naturalness. To this end, the reconstruction loss and bone length loss are utilized to regularize the learning model and train the network.

Reconstruction Loss: It aims to ensure the network model to preserve the information from the visible part of the motion sequence, and the Root Mean Squared Error (RMSE) [21] is utilized to measure the reconstruction loss:

$$\mathcal{L}_{rec} = ||\mathbf{M} \odot \mathbf{X}^{cor} + (1 - \mathbf{M}) \odot \mathbf{X}^{rec} - \mathbf{X}|| \tag{3.19}$$

Bone Length Loss: The bone length of kinematic model should be consistent across all frames. Therefore, we take the invariance property of the bone length to regularize the learning model:

$$\mathcal{L}_{bone} = \sum_{i=1}^{N} \sum_{j=1}^{J} ||l_{i,j}^{rec} - l_{i,j}||_2 \tag{3.20}$$

where the $l_{i,j}$ denotes the j-th bone length of i-th frame, $\mathcal{L}_{i,j}^{rec}$ is the corresponding recovered bone length. As a result, the total loss is the sum of reconstruction loss and bone length loss:

$$\mathcal{L}_{joint} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{bone} \tag{3.21}$$

where hyper-parameter λ is utilized to balance the contributions of two parts. It is noted that the missing joint is only reconstructed and the other parts are set as the equal value to the input data. Under such circumstances, some moving trajectories may be unsmooth between the recovered motion frames. To tackle this problem, we further utilize a sliding window to linearly smooth the local linearity of human motion trajectories. As a result, the smooth trajectories with spatio-temporal consistency can be well obtained with naturality and higher quality.

4. Experiments

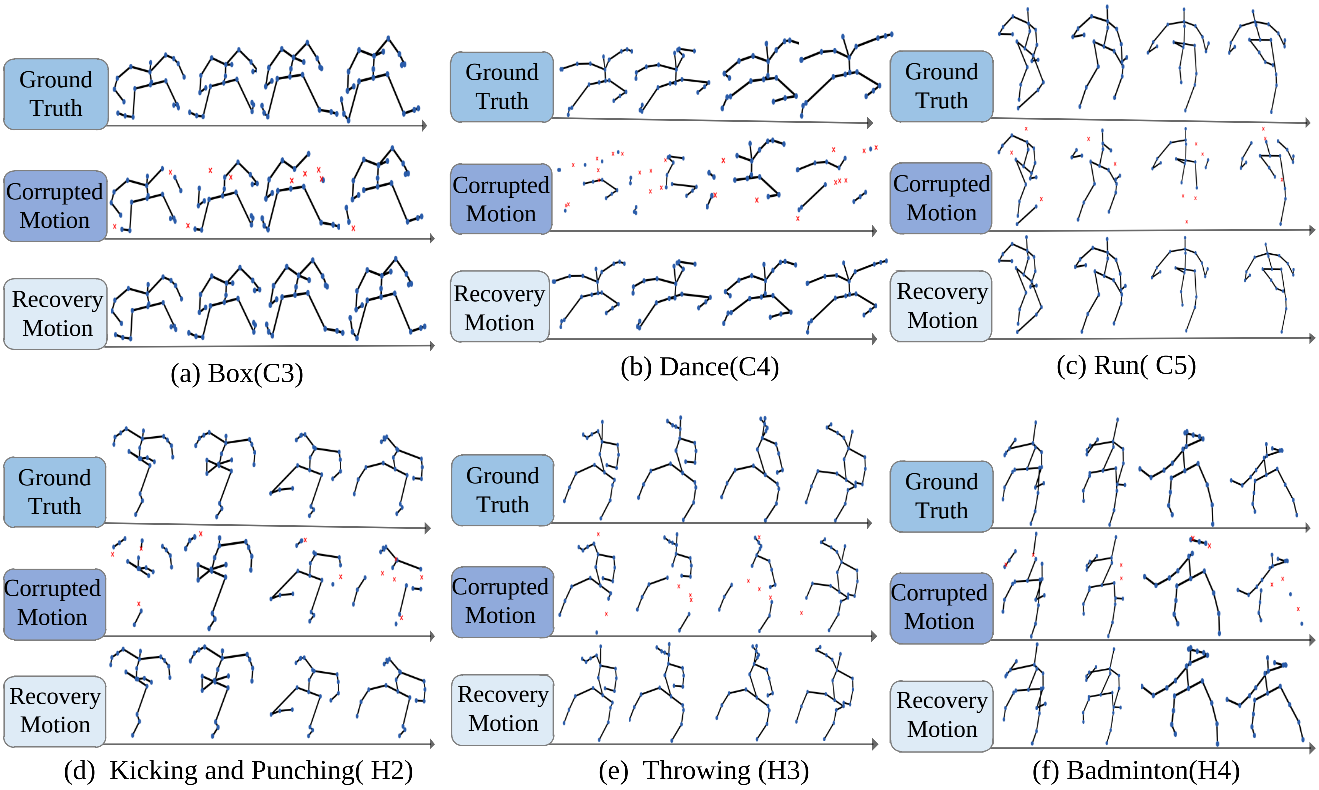
In this section, we conduct a series of quantitative experiments to evaluate the efficiency of the proposed motion recovery framework. The experimental results and comparison analysis will be detailed in the following subsection.

4.1. Dataset and evaluation metric

The public available CMU and HDM05 mocap datasets are selected for evaluation [14]. These datasets capture a large number of human actions with different semantics, such as walking, basketball, running, boxing and dance. Specifically, as shown in Table 1, ten representative sequences are chosen to evaluate the motion recovery performance. To simulate the random missing of motion frames in the sequence, we utilize the missing rate (MR) from 10% to 40% to control the proportion of missing joint entries, and randomly remove a certain number of active joints. Such processing ensures randomness in the position of missing joints and randomness in the position of missing frames in the motion sequence. In addition, we set K at 5 and fix the λ value to be 1.

Table 1. The statistical information about the tested mocap sequences.

Motion description (CMU)	Frames	Mark	Motion description (HDM05)	Frames	Mark
Swordplay	2240	C1	Dancing	8336	H1
Tai Chi	17,792	C2	Kicking and punching	6823	H2
Box	4800	C3	Throwing	3219	H3
Dancing	3136	C4	Badminton	2859	H4
Run	8384	C5	Clapping and waving	5607	H5



(a) Box(C3)

(b) Dance(C4)

(c) Run(C5)

(d) Kicking and Punching(H2)

(e) Throwing (H3)

(f) Badminton(H4)

Figure 3. Representative recovery frames selected from CMU and HDM05 databases.

The proposed RGN-TPR model is compared with the state-of-the-art competing methods, including long short-time memory network (LSTM) and fully connected neural network (FCNN) [5], bidirectional recurrent autoencoder (BRA) [22], attention-based LSTM network (A-LSTM) and least-squares (LS) constraint(A-LSTM+LS) [18]. Specifically, we utilize the same parameters as the authors have shared in their raw papers to recover the incomplete mocap sequences. To quantitatively evaluate the motion recovery performance, we utilize the reconstructed root mean square error (RMSE) to measure the difference between the reconstructed motion data and the original motion data. Meanwhile, the normalized position error, defined as the normalized 3D coordinate distance between the restored frame joints and ground truth, is utilized to evaluate the recovering performance.

4.2. Performance analysis and comparison

In the experiments, we randomly select some motion frames within significant missing joints for visual illustration, and representative incomplete motion recovering results are displayed in Figure 3. It

Table 2. Quantitative results obtained by different approaches under different missing rates.

Methods	BRA					FCNN					LSTM					A-LSTM+LS					Ours				
Motion	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
MR 10%	1.01	0.95	1.03	1.01	1.17	1.14	1.12	1.24	1.21	1.38	1.09	0.98	1.09	1.05	1.20	0.79	0.71	0.81	0.72	0.96	**0.42**	**0.31**	**0.41**	**0.33**	**0.51**
MR 20%	1.23	1.01	1.22	1.30	1.30	1.58	1.44	1.35	1.77	1.62	1.26	1.14	1.27	1.44	1.33	1.05	0.81	0.99	1.08	1.01	**0.57**	**0.40**	**0.51**	**0.42**	**0.52**
MR 30%	1.44	1.22	1.44	1.70	1.41	1.69	1.56	1.86	1.88	1.87	1.52	1.36	1.51	1.80	1.59	1.09	0.94	1.08	1.26	1.17	**0.58**	**0.43**	**0.52**	**0.48**	**0.55**
MR 40%	1.81	1.66	1.73	1.76	1.85	2.22	2.04	2.05	2.23	2.37	1.97	1.78	1.96	1.90	2.08	1.66	1.21	1.60	1.51	1.58	**0.63**	**0.44**	**0.54**	**0.51**	**0.59**
Methods	BRA					FCNN					LSTM					A-LSTM+LS					Ours				
Motion	H1	H2	H3	H4	H5	H1	H2	H3	H4	H5	H1	H2	H3	H4	H5	H1	H2	H3	H4	H5	H1	H2	H3	H4	H5
MR 10%	2.01	1.65	1.31	3.56	1.57	2.89	2.15	1.66	4.03	1.88	2.53	1.74	1.45	3.74	1.66	1.61	1.32	1.28	1.81	1.28	**0.52**	**0.62**	**0.49**	**0.66**	**0.60**
MR 20%	2.48	1.81	1.62	4.12	1.72	3.44	2.16	2.08	4.72	1.92	2.94	1.93	1.77	4.35	1.74	1.87	1.56	1.30	2.21	1.61	**1.14**	**1.14**	**0.89**	**1.02**	**1.02**
MR 30%	3.39	1.84	2.19	5.08	1.97	4.11	2.29	2.76	6.04	2.32	3.58	1.97	2.36	5.32	1.92	1.95	1.57	1.31	2.24	1.66	**1.36**	**1.45**	**1.07**	**1.35**	**1.39**
MR 40%	4.35	1.98	2.20	5.96	2.06	4.93	2.40	2.85	7.46	2.87	4.88	2.16	2.46	7.28	2.25	2.59	1.67	1.53	2.49	1.73	**1.47**	**1.45**	**1.10**	**1.45**	**1.40**

Table 3. Quantitative comparisons with different motion sequences.

Motion	C1				C3				C5			
	short-term		long-term		short-term		long-term		short-term		long-term	
missing time	500	1000	1500	2000	500	1000	1500	2000	500	1000	1500	2000
BRA	2.43	2.73	2.87	2.92	2.13	2.30	1.70	2.21	1.61	1.73	2.04	1.85
FCNN	2.92	2.96	3.15	3.19	2.42	2.98	2.64	2.42	1.71	1.86	2.12	2.94
LSTM	3.09	2.89	3.01	3.17	2.24	2.54	1.37	2.43	1.87	1.99	1.78	1.97
A-LSTM+LS	1.67	1.51	1.76	1.63	1.29	1.48	1.17	1.75	1.85	1.79	1.60	1.46
Ours	**1.56**	**1.48**	**1.51**	**1.45**	**1.11**	**1.32**	**1.01**	**1.67**	**1.74**	**1.66**	**1.54**	**1.37**
Motion	H1				H3				H4			
	short-term		long-term		short-term		long-term		short-term		long-term	
missing time	500	1000	1500	2000	500	1000	1500	2000	500	1000	1500	2000
BRA	3.49	4.28	4.63	3.98	1.14	1.90	2.11	2.20	4.71	5.72	5.96	5.34
FCNN	4.35	4.11	4.98	4.63	2.27	2.76	2.96	3.15	6.11	6.96	7.31	7.56
LSTM	3.98	4.15	4.78	4.99	2.19	2.40	2.66	2.54	7.24	7.40	7.72	7.74
A-LSTM+LS	1.87	1.53	1.80	1.71	1.15	1.57	1.54	1.98	1.91	2.14	2.40	2.48
Ours	**1.64**	**1.45**	**1.67**	**1.63**	**1.14**	**1.48**	**1.28**	**1.83**	**1.81**	**1.86**	**1.63**	**1.77**

can be well observed that the incomplete mocap collections generally fail to show a precise recordings of real human articulations. Although the missing markers are situated in a random and irregular way, the proposed motion recovery framework is able to well restore the missing entries. Visually, even the missing marker problem has significantly influenced the human poses, the completed motion frames almost exhibit the similar representations with the raw motion frames perceptually. Importantly, the recovered motions obtained by the designed RGN-TPR framework is able to well recover the real moving trajectories, and extensive experiments show its outstanding performance.

The motion recovery performances tested on different datasets and evaluated with different MR rates are list in Table 2, it can be found that the proposed RGN-TPR framework has delivered relatively lower RMSE scores on both CMU and HDM05 datasets, and these scores are always less than the results generated by the competing baselines. Under MR 30%, it can be found that the RMSE values obtained by the BRA, FCNN, LSTM and A-LSTM+LS methods are all higher than 0.1 when tested the sequences C1, C3, C4 and C5 on CMU dataset. It indicates that the recovered entries obtained by these competing baselines may not exactly match the real poses. A plausible reason is that the BRA

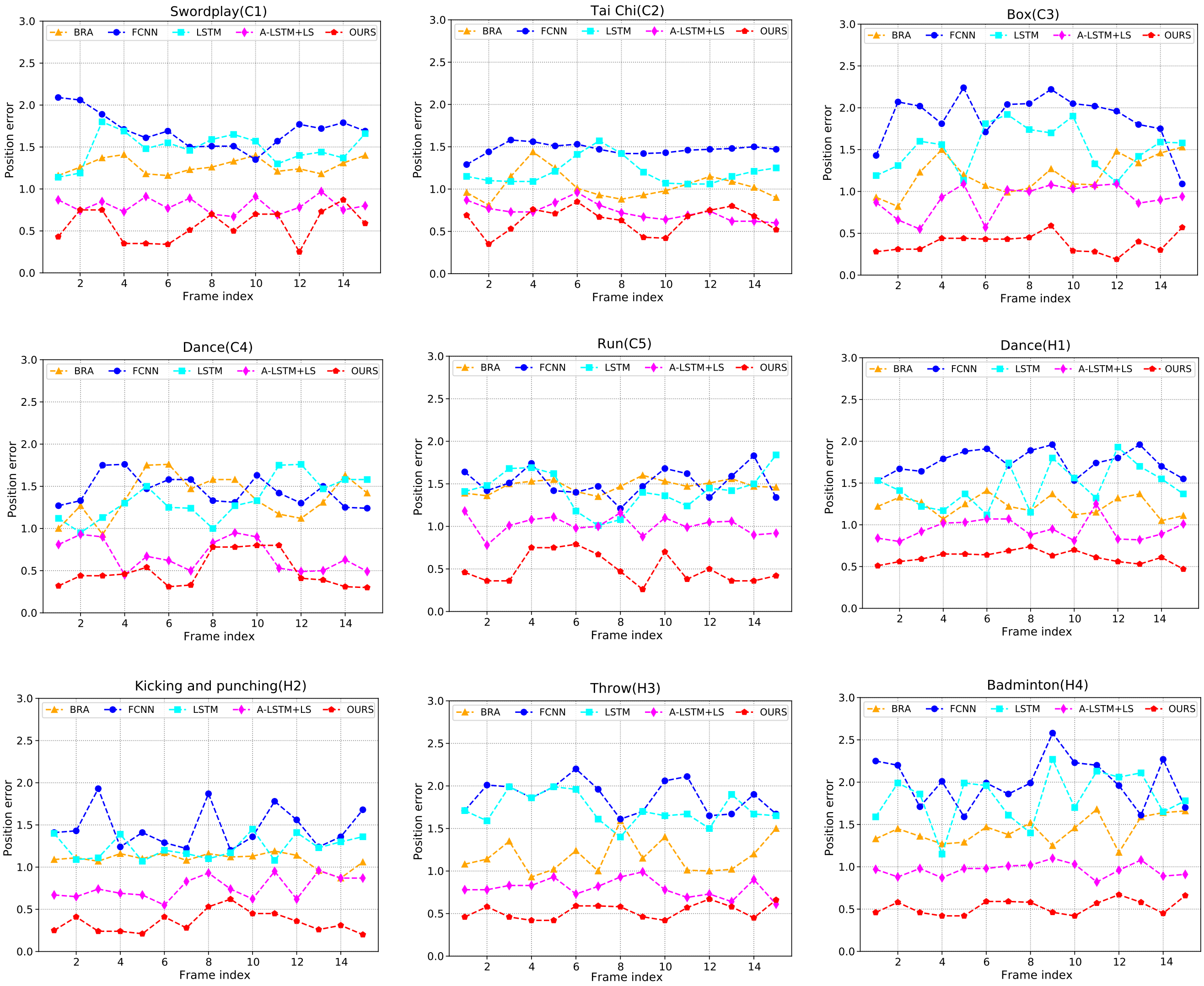


Figure 4. The normalized position errors obtained by different approaches and tested on different frames in the sequences.

and FCNN approaches are suitable to complete the missing motions with low complexity, but which often fail to investigate the articulated complexity and temporal property within the missing sequence. Accordingly, the restored performance may be very poor when the incomplete sequences incorporate with large spatio-temporal complexity. Although the LSTM and A-LSTM+LS methods attempt to boost the recovering performance by considering the dynamic information within the sequence, the RMSE scores obtained by these two approaches are also a bit large. The main reason lies that these two approaches feed the human skeleton directly into the deep networks to simulate the human motion model, which often ignore the local topological structure within the original data space and may fail to aggregate the structural relationships between different joints. As a result, some recovered frames may appear unnatural movements and deviate from the real motion data to a certain extent.

In contrast to this, the RMSE values obtained by the proposed approach are significant lower than the competing approaches. For different missing rates, the RMSE scores obtained by the proposed model are also less than the competing baselines. That is, the proposed RGN-TPR framework has yielded the best recovering performance. Although these incomplete sequences have the heterogeneous poses and different MR rates, the proposed RGN-TPR model is able to provide the reasonable estimations such that the restored results incorporate a precise representation to reveal the real movements. This indicates that the proposed learning model can effectively extract the spatio-temporal features hidden in the missing human skeleton structure and motion trajectories, which therefore can effectively reconstruct motion data in a reliable way.

In view of the continuous missing of adjacent markers in the same frame, we further perform a set of experiments to evaluate the recovery performance in the case of continuous missing. We select 6 sequences of 2000 frames in length and remove all information for specific joints (e.g., thigh, forearm) with 100 missing frame intervals. Table 3 displays the RMSE scores between the proposed RGN-TPR model and the competing baselines. It can be clearly observed that the performances of the existing methods drop sharply when the length of tested sequences is increasing. A possible reason lies that the spatio-temporal information of the motion sequence is severely damaged when the multiple markers in the sequence are continuously missing. Under such circumstances, it is difficult for the existing restoration methods to estimate the missing information from the available positions. By contrast, the proposed RGN-TPR model carefully considers the relationship-aggregated joint information, dynamic characteristics and long-term dependencies between the skeletons, which can capture the rich information from the available markers to complete the missing joints. In addition, we select the MR value to be 40% and further evaluate the normalized position error between the complete and restored motion sequences. As shown in Figure 4, it can be clearly observed that the position errors generated by the proposed RGN-TPR method are always less than the error results obtained by the competing approaches. Note that, the small normalized position error indicates the better recovering performance, and the recovered results are able to well match the real movements. It can be observed that the incomplete motions recovered by the proposed RGN-TPR approach are able to well match the real movements. Even a large proportion of markers is missing or the markers are missing for a long period of time, the proposed RGN-TPR framework has yielded the best motion recovering performances, and the completed missing entries are able to well match the real movements.

5. Conclusions

This paper presents an efficient mocap data recovery approach by using relationship-aggregated graph network and temporal pattern reasoning mechanism. The proposed framework utilizes local graph encoder to encode the high-level semantic node features and their semantic relationships in each local part, and employs the global graph encoder to aggregate the structural relationships between different parts for whole skeletal data representation. Meanwhile, the designed temporal pattern reasoning mechanism is able to exploit the intra-frame interactions and capture long-term dependencies between the motion frames, whereby the discriminative spatio-temporal features and semantic correlations can be reasonably obtained for efficient motion recovery. Extensive experiments conducted on various kinds of motion sequences have shown its outstanding performance.

Acknowledgments

This work was supported in part by the National Science Foundation of Fujian Province under Grant 2020J01083, in part by the Program for Excellent Young Talents in the University of Fujian Province under Grant JAT191922 and Grant JAT190847, in part by the Xiamen Medical College under Grant JT180642, Grant XBJG2018002 and Grant FBJG20200368.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. P. Zubova, P. Svetlana, K. A. Raetskiy, Modeling the trajectory of motion of a linear dynamic system with multi-point conditions, *Math. Biosci. Eng.*, **29** (2021), 7861–7876. https://doi.org/10.3934/mbe.2021390
2. J. H. Zhou, S. Jia, J. B. Chen, M. Chen, Motion and trajectory planning modeling for mobile landing mechanism systems based on improved genetic algorithm, *Math. Biosci. Eng.*, **18** (2021), 231–252. https://doi.org/10.3934/mbe.2021012
3. X. Liu, G. F. He, S. J. Peng, Y. M. Cheung, Y. Y. Tang, Efficient human motion retrieval via temporal adjacent bag of words and discriminative neighborhood preserving dictionary learning, *IEEE Trans. Hum.-Mach. Syst.*, **46** (2021), 763–776. https://doi.org/ 10.1109/THMS.2017.2675959
4. D. Holden, J. Saito, T. Komura, A deep learning framework for character motion synthesis and editing, *ACM Trans. Graphics*, **35** (2016), 1–11. https://doi.org/10.1145/2897824.2925975
5. Q. Cui, H. Sun, Y. Li, Y. Kong, A deep bi-directional attention network for human motion recovery, in *Proceedings of the International Joint Conference on Artificial Intelligence*, (2019), 701–707. https://doi.org/10.5555/3367032.3367133
6. Z. Yang, Y. D. Yan, H. T. Gan, J. Zhao, Z. W. Ye, A safe semi-supervised graph convolution network, *Math. Biosci. Eng.*, **19** (2022), 12677–12692. https://doi.org/10.3934/mbe.2022592
7. H. Yuan, J. Huang, J. Li, Protein-ligand binding affinity prediction model based on graph attention network, *Math. Biosci. Eng.*, **8** (2021), 9148–9162. https:/doi.org/10.3934/mbe.2021451
8. A. Aristidou, J. Cameron, J. Lasenby, Real-time estimation of missing markers in human motion capture, in *Proceedings of International Conference on Bioinformatics and Biomedical Engineering*, (2008), 1343–1346. https:/doi.org/10.1109/ICBBE.2008.665
9. Q. Wu, P. Boulanger, Real-time estimation of missing markers for reconstruction of human motion, in *Proceedings of Symposium on Virtual Reality*, (2011), 161–168. https:/doi.org/10.1109/SVR.2011.35
10. M. Burke, J. Lasenby, Estimating missing marker positions using low dimensional kalman smoothing, *J. Biomech.*, **49** (2016), 1854–1858. https://doi.org/10.1016/j.jbiomech.2016.04.016
11. R. Y. Lai, P. C. Yuen, K. Lee, Motion capture data completion and denoising by singular value thresholding, in *Proceedings of IEEE international conference on eurographics*, (2011), 45–48. http://dx.doi.org/10.2312/EG2011/short/045-048

12. C. H. Tan, J. Hou, L. P. Chau, Human motion capture data recovery using trajectory-based matrix completion, *Electron. Lett.*, **49** (2013), 752–754. https://doi.org/10.1049/el.2013.0442

13. Y. Feng, J. Xiao, Y. Zhuang, X. Yang, J. J. Zhang, R. Song, Exploiting temporal stability and low-rank structure for motion capture data refinement, *Inf. Sci.*, **277** (2014), 777–793. https://doi.org/10.1016/j.ins.2014.03.013

14. S. Peng, G. He, X. Liu, H. Wang, B. Zhong, Motion segmentation based human motion capture data recovery via sparse and low-rank decomposition, *J. Comput.-Aided Des. Comput. Graphics*, **27** (2015), 721–730.

15. L. Herda, P. Fua, R. Plankers, R. Boulic, D. Thalmann, Skeleton-based motion capture for robust reconstruction of human motion, in *Proceedings Computer Animation*, (2000), 77–83. https://doi.org/10.1109/CA.2000.889046

16. L. Li, J. McCann, N. S. Pollard, C. Faloutsos, Dynammo: Mining and summarization of coevolving sequences with missing values, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2009), 507–516. https://doi.org/10.1145/1557019.1557078

17. J. Xiao, Y. Feng, W. Hu, Predicting missing markers in human motion capture using l1-sparse representation, *Comput. Anim. Virtual Worlds*, **22** (2011), 221–228. https://doi.org/10.1002/cav.413

18. L. Ji, R. Liu, D. Zhou, Q. Zhang, X. Wei, Missing data recovery for human mocap data based on a-lstm and ls constraint, in *Proceedings of IEEE International Conference on Signal and Image Processing*, (2020), 729–734. https://doi.org/10.1109/ICSIP49896.2020.9339359

19. C. Xie, J. Lv, Y. Li, Y. Sang, Cross-correlation conditional restricted boltzmann machines for modeling motion style, *Knowledge-Based Syst.*, **159** (2018), 259–269. https://doi.org/10.1016/j.knosys.2018.06.026

20. S. K. Tian, N. Dai, L. L. Li, W. W. Li, Y. C. Sun, X. S. Cheng, Three-dimensional mandibular motion trajectory-tracking system based on BP neural network, *Math. Biosci. Eng.*, **17** (2020), 5709–5726. https://doi.org/10.3934/mbe.2020307

21. X. Liu, Z. Hu, H. Ling, Y. M. Cheung, MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 964–981. https://doi.org/10.1109/TPAMI.2019.2940446

22. S. Li, Y. Zhou, H. Zhu, W. Xie, Y. Zhao, X. Liu, Bidirectional recurrent autoencoder for 3d skeleton motion data refinement, *Comput. Graphics*, **81** (2019), 92–103. https://doi.org/10.1016/j.cag.2019.03.010