Attention-Based Recurrent Autoencoder for Motion Capture Denoising

Yongqiong Zhu[1], Fan Zhang[2], Zhidong Xiao[3*]

[1] School of Art, Wuhan Business University, China
[2] Mathematics and Computer Science School, Wuhan Polytechnic University, China
[3] National Centre for Computer Animation, Bournemouth University, UK
yongqiongzhu@163.com, whpuzf@whpu.edu.cn, zxiao@bournemouth.ac.uk

Abstract

To resolve the problem of massive loss of MoCap data from optical motion capture, we propose a novel network architecture based on attention mechanism and recurrent network. Its advantage is that the use of encoder-decoder enables automatic human motion manifold learning, capturing the hidden spatial-temporal relationships in motion sequences. In addition, by using the multi-head attention mechanism, it is possible to identify the most relevant corrupted frames with specific position information to recovery the missing markers, which can lead to more accurate motion reconstruction. Simulation experiments demonstrate that the network model we proposed can effectively handle the large-scale missing markers problem with better robustness, smaller errors and more natural recovered motion sequence compared to the reference method.

Keywords: Motion capture, Attention mechanism, Deep learning, Neural network

1 Introduction

Motion capture is used to record the movement of actors and then retarget the recorded data to CG characters for animation, and has been used commercially in games, film effects and virtual reality. Currently, optical motion capture systems remain the dominant technology, using multiple cameras to track markers affixed to the actor’s body and, finally, to reconstruct the individual’s 3D marker trajectory from a 2D image by means of triangulation techniques that offer high precision, flexibility and ease of use. However, the raw motion sequence obtained directly from optical capture devices is often corrupted or incomplete, which can occur for a number of reasons such as calibration errors, poor sensor resolution, body movement or occlusion caused by costumes. As a result, a great deal of effort is often required to clean the data before it can be used for motion retargeting. Although most commercial equipment offers a software system that intelligently handles noise, much of the process requires manual human intervention and does not work satisfactorily under different motions, such as a high percentage of missing markers or occlusion.

The optical motion capture system captures the three-dimensional position of each marker over time. When occlusion or other conditions occur, the capture system loses track of the marker for a period of time, resulting in a gap. So, the challenges to denoising are: 1. markers may be lost at any time during the capture process and the missing features do not satisfy a specific distribution; 2. when markers are lost at long gaps, it is difficult to obtain useful information from contextual that is relevant to corrupt data recovery; 3. the recovered motion often behave unnaturally. These factors present significant challenges to effective motion recovery.

Most commercial motion capture systems provide a variety of software to edit MoCap data for later denoising and missing padding. Most of the algorithms used by these softwires are based on linear and non-linear interpolation algorithms, which are only effective when dealing with small scale missing data, and are less effective when processing data containing large scale corrupted data, resulting in distorted and unnatural motion sequences after recovered, which still require manual adjustment by the animator. In the early days of research, signal processing-based methods were a very important approach [1-2]. Considered from the point of view of signal filtering, different MoCap data filters were constructed. Standard Gaussian filters, kinetic filters and Kalman filters were all applied for denoise processing. The signal processing approach only works well with small-scale corrupted data. Later, Lai [3] first noticed that the motion sequence matrix has the characteristic of low rank, so the missing noise problem of the MOCAP sequence can be transformed into a matrix low rank recovery problem and the proposed objective function can be solved by the singular soft thresholding method. The greatest advantage of this type of algorithm is that it solves the missing sample problem. However, none of these algorithms make use of the structural information implicit in the human motion capture sequence.

In recent years, with the development of technology, motion capture has developed rapidly, and many free MoCap data have emerged for users to use, providing enough samples for the research of denoising. And with the widespread use of deep learning algorithms, researchers also found that deep learning-based neural networks are particularly suitable for the refinement of MoCap data. Fully connected neural networks, recurrent neural networks LSTM, residual neural networks ResNet have all been used to predict missing marker data. These neural networks are excellent at processing sequence data and can effectively exploit the temporal and spatial information hidden in the MoCap data, but have the possibility to produce “gradient disappearance” and “gradient explosion”. Later, Holden [4] proposed a neural network model with an automatic encoder for motion sequence based on manifold learning theory, constructing a multi-layer fully connected

*Corresponding Author: Zhidong Xiao; E-mail: zxiao@bournemouth.ac.uk

DOI: 10.53106/160792642022112306015

layer of decoder to project motion sequences to learn attributes of sequences and perform dimensional conversion, and then inverse projecting them through a decoder, which has greatly improved the recovery accuracy for missing noise and jitter noise, similar network models are ERD [5], BRA [6], etc. Although these models had made practical progress in long motion sequences, the prediction efficiency can be reduced because of the difficulty of capturing long-term time dependence in the recurrent unit. In addition, different frame positions should contribute unequally to denoising, and the previously proposed approaches does not differentiate from the context based on contribution values, which may also lead to a reduction in prediction efficiency.

Recently, in deep learning for sequence data, attention mechanisms [7] have been widely used in machine translation, natural language processing, and image processing because of their effectiveness in capturing the long-term temporal dependence of sequences. The attention mechanism is embedded in the recurrent network model used to process sequential data, where different weights value is measured from the context to select relevant data as input during the encoding or decoding phase, rather than treating all contexts equally. Different frames in a motion sequence should have different weight values, and theoretically, a reasonable selection of attribute information associated with a corrupted frame with a greater weight can improve recovery accuracy. Therefore, introducing the attention mechanism into the refinement of motion capture should be a feasible approach.

Therefore, inspired by the theory of manifold learning and attention mechanism, this paper proposes an automatic encoder-decoder neural network architecture. The advantage of the proposed method is that the use of the encoder enables automatic human motion manifold learning to capture the hidden spatial-temporal relationships in motion sequences, this greatly improves the accuracy of denoising. Moreover, more accurate recovery can be obtained by repairing the damaged motion through an attention mechanism that finds the specific positional information that are most correlated with corrupted frames. Finally, in the process of motion reconstruction, the influence of marker position and bone length on motion sequence reconstruction is also considered, so that the restored motion sequence is more natural and can reflect the real motion more effectively.

The main work of this paper is as follows:

1. A neural network motion capture denoising architecture is proposed to extract the manifold learning of human motion through an encoder-decoder with a Bi-LSTM. The model can effectively exploit spatial-temporal features from the movement sequence.

2. The model adds an attention mechanism, which can effectively obtain the long-term feature information of the recurrent neural network, and pay attention to the most important position information by calculating the attention weight value.

3. The effect of errors in bone length on the kinematic reconstruction is taken into account in the loss function, which makes the reconstructed motion sequence more natural and smoother, and more accurately reflects the true motion.

The remaining part of the paper proceeds as follows: In the second part, related work on denoising methods and attention mechanisms are investigated. In the third section, we proposed our network model. In the fourth section, extensive experiments are performed to evaluate the proposed method, and the conclusions of the paper are given in the fifth section.

2 Related Work

2.1 Denoising Method

Research on motion capture data denoising is usually divided into two categories: non-data-driven based approaches and data-driven based approaches.

1. Non-data-driven based approaches

The main non-data-driven methods are the interpolation methods, the filter methods and the matrix methods.

The interpolation method mainly uses adjacent markers to infer missing data and is used by most commercial software (e.g. Vicon) due to its simplicity and effectiveness. Howarth [8] compared three different interpolation techniques: linear interpolation, cubic spline interpolation and local coordinate system (LCS) interpolation. Gløersen [1] solved the problem of small-scale lossing markers, and the performance is better than the traditional spline interpolation method, however, the results are not satisfactory when a large number of markers are missing.

The filter approach is based on kinetic simulation algorithms that transform physically uncoordinated motion in a motion sequence into coordinated motion. Shin [9] used the Kalman filter to deliver human movement to a computer character in real time for the first time. Hsieh [10] based on wavelet theory to remove impulse noise from motion sequences. He used multiresolution analysis to decompose noise data, the noise is assigned as coefficients of high magnitude, and then denoise them by smoothing these high-magnitude coefficients.

If a human motion sequence is represented as a matrix (rows indicate number of frames, columns indicate coordinates), the matrix have low rank characteristics. A low-rank matrix is a form of sparse representation, i.e. using a matrix of lower rank to approximate the original matrix not only preserves the main features of the original matrix, but also reduces the storage space and computational complexity of the data. Based on this theory, Lai [3] used reduced rank of the matrix to solve denoising problem of motion capture by filling the matrix with a singular value shrinkage algorithm (SVT). This method proved to be effective in exploiting the low-rank property, but only for the specified motion. Xiao [11] subdivided the motion poses into multiple partitions for each human pose, and then processed these partitions separately. However, if markers are missing making an entire row or column in the matrix loss, it is difficult to reconstruct and cannot repair severely corrupted human motion data.

Non-data-driven methods are fast in denoising, but can only handle a small range of missing markers. In addition, none of these algorithms make good use of the temporal and spatial in context implicit in motion sequences and cannot be used to handling large scale data loss problems.

2. Data-driven based approaches

In recent years, data-driven approaches that train deep learning models on large-scale data and then perform noise prediction have been applied to many studies in computer graphics, for example, image restoration and image denoising [12-13]. Inspired by these studies, many scholars have applied deep neural networks to denoising motion capture data, which has gradually become a mainstream technology.

In 2015, Holden [4] present a technique for learning a manifold of human motion data using Convolutional Autoencoders. Since the position of a single marker is not only related to the positions of other markers in the same frame, but also related to the positions of markers in other frames in the motion sequence, the manifold can be treated as the prior probability distribution of motion data, and large-scale training focused on noise data can be performed. However, using a convolutional network as an encoder is prone to jitter after motion reconstruction. In the same year, Fragkiadaki proposed an encoder-decoder-based recurrent neural network (ERD) network [5]. Compared with aperiodic motion, the ERD model performs better on periodic motion, but the restoration results are not naturally. In 2017, Mall proposed the EBF model [14] based on the ERD model, which requires the skeletal structure of the human body to be constructed using the EBD model first and then cleanup using the EBF model. Obviously, building a skeleton from all the frames takes a lot of time. In 2018, Kucherenko [15] proposed the use of LSTM networks with sliding windows to alleviate the long-distance information learning problem of recurrent neural networks. In 2019, Li constructed a bidirectional LSTM network [6] to deal with jitter noise in motion sequences and achieved good results.

The data-driven approach relies on a large amount of a priori data and uses deep learning technology to learn the spatial-temporal relationships of sequence data in a low-dimensional feature space, which has better universality and robustness. However, the current data-driven approach based on motion reconstruction is not very effective, mainly because when learning in the context, it does not effectively identify which information is important to contribution weight value. As they consider the importance to be equal, the learning effect is not very effective. In addition, the recovered movement is often less natural and there is a certain amount of jitter.

2.2 Attention Mechanism

The attention mechanism is a means by which humans use their limited attentional resources to quickly filter out high-value information from a large amount of information. The human visual attention mechanism has greatly improved the efficiency and accuracy of visual information processing. In essence, the attention mechanism in deep learning is similar to the human selective visual attention mechanism, and the core goal is to select the information that is more critical to the current task goal from a large amount of information, which is the core idea of the Attention Model in deep learning. Therefore, AM was initially applied in the area of images. Later, the researchers applied the AM model to the NLP field and proposed the Soft Attention Model [16]. In 2017, Vaswani [17] proposed a new network architecture, namely Transformer, which uses a multi-head attention mechanism to connect the encoder and decoder, and achieved better results in machine translation, enabling the attention mechanism become the hot of research, researchers have proposed many attention models, such as SE [18] and PFA [19].

Researchers soon applied the attention mechanism to motion capture. In 2020, Cui proposed the BAN network model [20]. He considers that current single-directional recurrent networks are insufficient for motion recovery, since a corrupted frame should inherently correlate with forward-propagation and back-propagation. Therefore, he uses Bi-LSTM and attention mechanism to simultaneously capture long-term temporal dependencies of sequences from forward and backward. Similar study includes the CRNN network model [21]. However, compared with the traditional methods, the multi-head attention mechanism can extract the feature information hidden in the context more effectively, has strong parallel computing and works well to prevent overfitting. Therefore, we intend to use a multi-head attention mechanism combined with recurrent neural network for effective denoising.

3 Methodology

This section describes the problem and presents the proposed denoising model. Then, the workflow of the network model is discussed and finally its training process is given.

3.1 Problem Description

The process of denoising based on deep learning network is shown as Figure 1. First, all the motion sequence data need to be pre-processed and normalized to get the ground truth, add noise to ground truth to obtain corrupted dataset, and set the training dataset and test dataset. Then the neural network architecture is designed, input the training dataset to the neural network to calculate the loss L, and then obtain the optimal network parameters through back propagation. Finally, the test dataset motion sequences are input and the predicted noise locations are repaired to obtain the reconstructed motion sequences. The goal of denoising is to make the loss as small as possible so that the reconstructed motion frame sequence can restore the original motion sequence as realistically as possible.

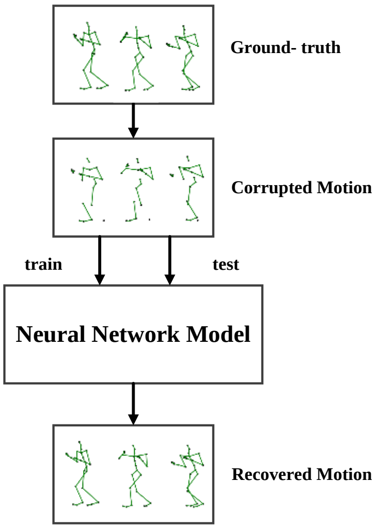


Figure 1. Refinement pipeline

Suppose a human motion capture sequence $X = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$ has n frames, and for any frame x_i, where $1 \leq i \leq n$ is the number of markers attached to the human body. Assume that the original ground-truth sequence is x^G, and the denoised sequence is Y, then the loss L:

$$L = f(Y - X^G). \quad \textbf{(1)}$$

The goal of denoising is to find this function f, such that the value of L is minimized.

3.2 Network Architecture

In this section, MAB architecture is proposed as Figure 2, a novel multi-head attention-based bidirectional recurrent encoder-decoder network architecture for denoising human motion data. MAB has three main features: (1) Extraction of human motion manifold learning by constructing encoders and decoders network model; (2) a multi-head attention method is used to obtain the most relevant frame position information in the context to obtain more accurate noise recovery; (3) Kinematic information is taken into account and the addition of bone length as a measure of the loss function makes the recovered sequence movement more natural.

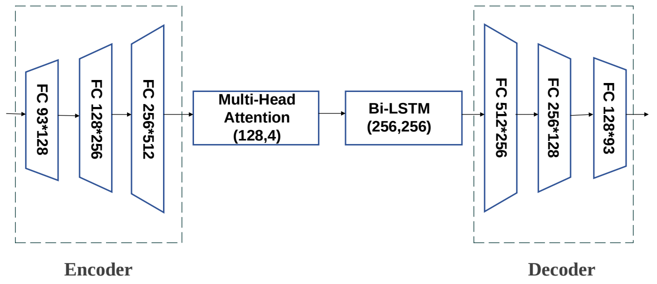


Figure 2. MAB architecture

MAB consists of four components: an encoder, an attention layer, a recurrent layer and a decoder. The encoder, composed of three fully connected networks with different neural units, transforms the input motion sequence in dimension to complete the projection and obtains the structural information implied in it through manifold learning. The motion sequences are transformed into a 512-dimensional vector matrix after passing through the encoder. The output of the encoder is used as input to the multi-headed attention layer, which uses four heads, each with 128 dimensions. The attention layer outputs a 512-dimensional vector and then passes it through a bidirectional LSTM with 256 hidden units to obtain a vector matrix of size n x 512, which is then fed to the decoder. As the decoder and encoder have a symmetrical framework, the output of the recurrent layer is inverse projected through the decoder to obtain the motion sequence.

3.3 Attention Mechanism

The attention mechanism used in this article is multi-head attention mechanism refers to the network model of [18]. The output of the encoder is linearly transformed and multiplied by the network parameters W^Q, W^K, W^V to get the input vectors Q, K, and V of the attention layer. Then, according to the number of heads used, the obtained Q, K, and V are further divided into h parts.

For each head $head_i$, calculate its attention value as:

$$head_i = Attention(Q_i, K_i, V_i) = soft\,max(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i, \quad \textbf{(2)}$$

where d_k denotes the dimension of the key of each head.

The results of multiple heads are connected and multiplied by the weight matrix parameter W^O to get the output of the attention mechanism.

$$MultiHead(Q, K, V) = \\ Concat(head_1, head_2, \cdots, head_h)W^O. \quad \textbf{(3)}$$

3.4 Evaluation of Loss

In the training process, the loss is generally measured by calculating the offset of the coordinates of the markers of the motion repair sequence from the original coordinates, defined as the position loss $Loss_p$. The specific calculation procedure is as follows.

$$Loss_p = \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{X}_1)^2, \quad \textbf{(4)}$$

where $\tilde{X}_1$ is the denoised frame.

However, experiments in [6] show that if the loss function is judged only by the loss of position recovered from the markers, the recovered motion sequence may still be jittered and unnaturally. So, the loss function cannot consider only the difference in marker positions. The presence of bone length jitter between frames can lead to unsmooth motion sequences. Therefore, it is difficult to produce the ideal restoration results using only marker position loss. The solution is to directly impose a bone length constraint in addition to the position constraint during the training phase of the network, so that the network output data has a more stable and reasonable bone length between frames, which is defined as $Loss_b$:

$$Loss_b = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J-1}\left|\left\|p_j^{i(x,y,z)} - p_{j+1}^{i(x,y,z)}\right\|_2 - \left\|\tilde{p}_j^{i(x,y,z)} - \tilde{p}_{j+1}^{i(x,y,z)}\right\|_2\right|. \quad \textbf{(5)}$$

Among them, the superscript of the variable p represents the three-dimensional coordinate corresponding to the point p in the original frame i, and the subscript represents the bone number, which is composed of two joint points j and $j + 1$, and $\tilde{p}$ is the restored frame. The final loss function is:

$$L = \lambda Loss_p + (1 - \lambda)Loss_b, \quad \textbf{(6)}$$

and $\lambda \in (0,1)$.

4 Simulation

We first introduce the human motion capture database, data preprocessing scheme, and model parameters used in the simulation. Then, we select three state-of-the-art human motion recovery approaches as reference, describe their implementation details and give experimental metrics for qualitative and quantitative analysis. Finally, the experiments were conducted to compare the performance of our method with the reference methods at different levels of markers missing ratios and at different missing time gaps.

4.1 Experimental Design

We use the CMU motion capture database, which is the most used free database for motion capture. It contains 2440 motion sequences in 111 subject categories, which provides good simulation conditions. we choose the ASF/AMC data format which can represent a motion sequence that contains some semantics, such as a running or dancing.

The AMC file gives information on the coordinates of the markers, but these are their coordinates in the local coordinate system of their parents and need to be converted from local to world coordinates by iterative transformations such as rotations and translations. We assumed that the markers are the locations of the joints, so there is a total of 31 markers on the human body. This experiment evaluates the performance of our method by randomly selecting action data from the CMU database for four subjects: playing basketball, walking, boxing and everyday behavioral actions.

1. Data preprocessing

Before inputting the motion sequence into the neural network, the data needs to be pre-processed to accelerate the convergence of the model. The procedure is as follows.

(1) Scaling

Before training the model, scale all marker coordinates in the sequence so that the motion capture actor has a uniform height. This normalization allows the motion capture system to not have to deal with character objects of different heights individually, and scale them with a scale factor. The scale factor can be calculated from the average length of the character's skeleton; it can also be extracted from the capture system during the calibration phase. Besides, the method in this paper does not make any assumptions about capturing topics, and can handle a variety of different types of topics well, ensuring that it has good generalization performance.

(2) Unified the coordinate system

The pose for each frame of the motion sequence needs to be converted into a world coordinate system with the root as the origin, and the horizontal direction from the character's left shoulder to the right shoulder is identified as the x-axis using the y-axis direction in the world coordinate system as the reference, and the z-axis is generated based on the normal vector between the x-axis and the y-axis.

(3) Normalization

The average position of the motion sequence is obtained, the marker positions are transformed with the average position as the center, and finally the marker positions are uniformly scaled to between [-1, 1] to obtain the dataset.

2. Generation of corrupt motion and datasets

To simulate noise in a motion sequence, we use a binary bit vector to indicate whether a marker is usable or occluded, with value 1 indicating that the marker is live and value 0 indicating that the marker is lost during the capture. For example, lwrist represents the joint on the left wrist, and a value of 1 means that the movement of the marker can be captured normally. If the marker is occluded when capture, the value of its bit vector is zero. To simulate the situation of missing markers, a vector mask of $\{0, 1\}$ needs to be randomly generated with Bernoulli distribution. A certain number of joints are randomly removed by dot product between vectors, which is consistent with the random loss of joints.

We add noise to the dataset obtained after pre-processing to form the dataset for the training of the network model. In the construction of the dataset, time_step is set to 5. Then 60% of the dataset is used as the training dataset, 20% of the data as the validation dataset and the remaining as the test dataset.

3. Network train parameters

The neural network model training parameters include the number of neural units in the input, hidden and output layers of the neural network, the setting of parameters such as network batch, dropout, epoch, and the execution of forward propagation algorithms, objective optimization, and backward propagation algorithms after completing the training of the neural network to determine the network parameters W, V and offset b.

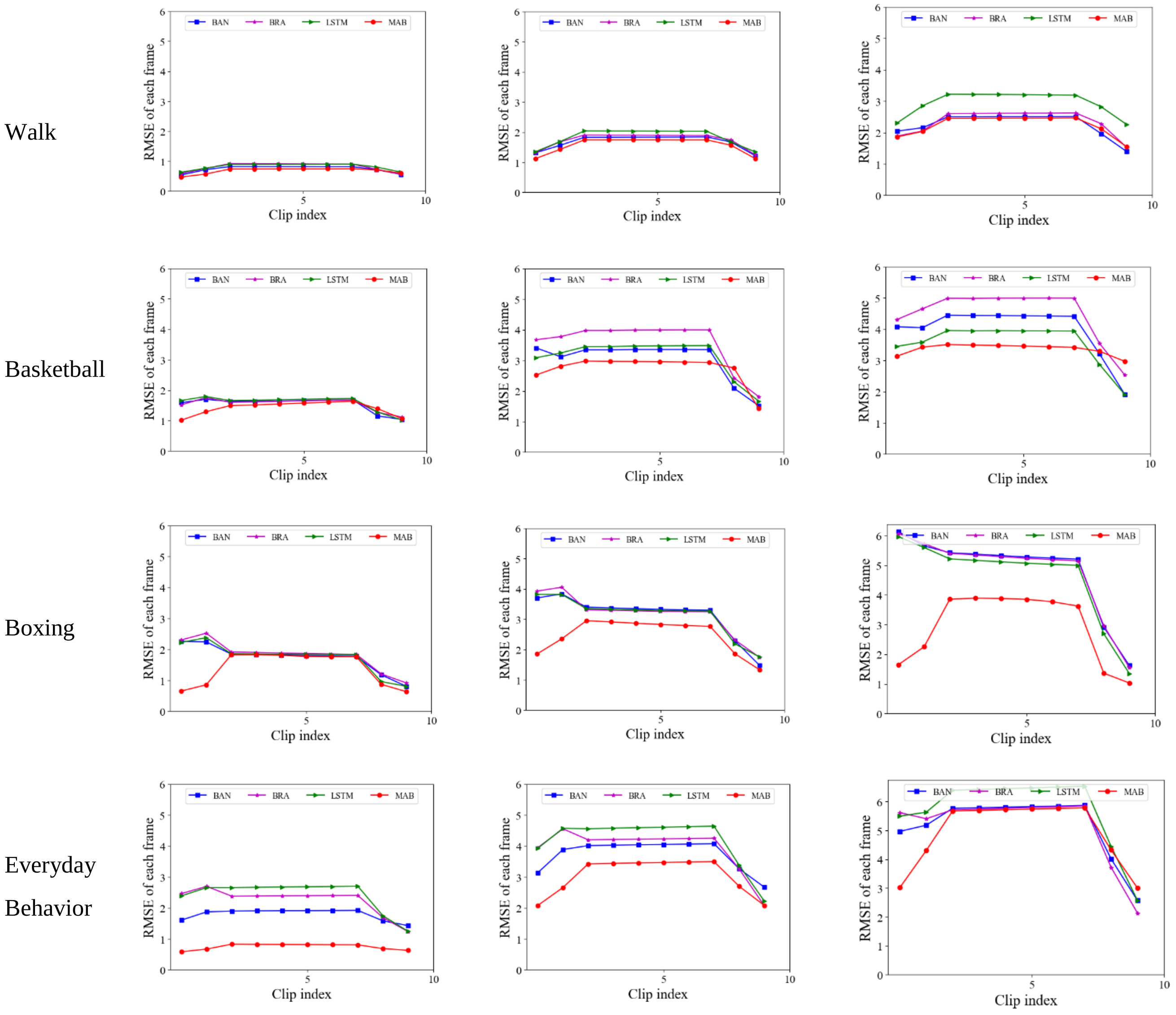
In the MAB network architecture proposed in this paper, the encoder and decoder use a multi-layer fully connected network to capture the forward and backward bi-directional dependencies of the damaged frames respectively, enabling better manifold learning of motion sequences and avoiding over-fitting. In the multi-head attention layer, h with a value of 4, and the input and output dimensions of each head are 128. The bidirectional LSTM hidden layer unit is 512 dimensions. In addition, we use dropout=0.4 as the regularization method on the LSTM layer to make the model have better generalization ability. $\lambda = 0.95$ set in the loss function.

The entire network model is randomly initialized with the weight and offset parameters, the batch size is 96, the epochs are 100 and our model is optimized throughout the training using the Adam function, the learning rate is 0.001. For each type of sequence, we trained the network a total of 10 times and took the average value for comparison. The implementation of all methods is based on the tensorflow2.0 framework and is performed with i7-12700K CPU, 3070Ti GPU and 64GB RAM.

In order to better compare the performance of the proposed model, three state-of-the-art methods were chosen as benchmark methods, that are BRA [6], LSTM [15] and BAN [20]. BRA is an BLSTM structure without using an attention mechanism. The LSTM [15] are all standard recursive units. BAN uses a general attention mechanism. Experimental indicators were evaluated using RMSE (reconstructed mean squared error) and BLE (Bone length error), in line with the benchmark methods.

4.2 Simulation Results and Analysis

In this section, we simulate various markers missing in a real capture procedure to quantitatively evaluate the accuracy of our method as well as the reference methods. The above model was evaluated under the same setting of random missing markers. We randomly delete a specific number of markers (10%, 30% and 50%) over several timeframes and apply each method to restore them. The gap length to markers missing is set to 10 frames. Figure 3. presents a comparison of the RMSE values of several motion sequences after different motion reconstruction methods.



(a) 10% (b) 30% (c) 50%

Figure 3. Comparison of RMSE of our method and the reference methods on four motion sequences at different level of markers missing ratio

In Figure 3, the first column is the RMSE comparison with 10% random missing points, and the second and third columns are RMSE comparisons with 30% and 50% missing points, respectively. We can see from Figure 3, its RMSE values are relatively low compared to regular sports (walk) and irregular sports (basketball, daily behavior, boxing). But no matter what kind of motion, the RMSE value of each model is increasing with the greater the loss rate of marker points, but compared with other methods, MAB can obtain lower RMSE value. Although both MAB and BAN use an attention mechanism, MAB is able to obtain more relevant location information in the contextual spatial-temporal relationships to recover the motion as it uses a multi-headed attention mechanism that can focus on more attribute to calculate the attention weights value compared to BAN. The specific RMSE values for each motion at different missing ratio are given in Table 1.

Table 1. Quantitative comparisons of RMSE of our method and the reference methods on four motion sequences at different level of markers missing ratio

Motion	Walking			Basketball			Boxing			Everyday behavior		
Missing ratio	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
OURS	0.68	1.56	2.23	1.02	2.82	3.14	1.24	2.45	2.99	0.74	2.57	4.9
BAN	0.74	1.67	2.26	1.59	3.02	4.08	1.73	3.13	4.82	1.79	3.56	5.15
BRA	0.811	1.73	2.34	1.52	3.56	4.31	1.82	3.16	4.79	2.24	3.92	5.17
LSTM	0.823	1.82	2.94	1.66	3.11	3.45	1.74	3.14	4.62	2.41	4.17	5.69

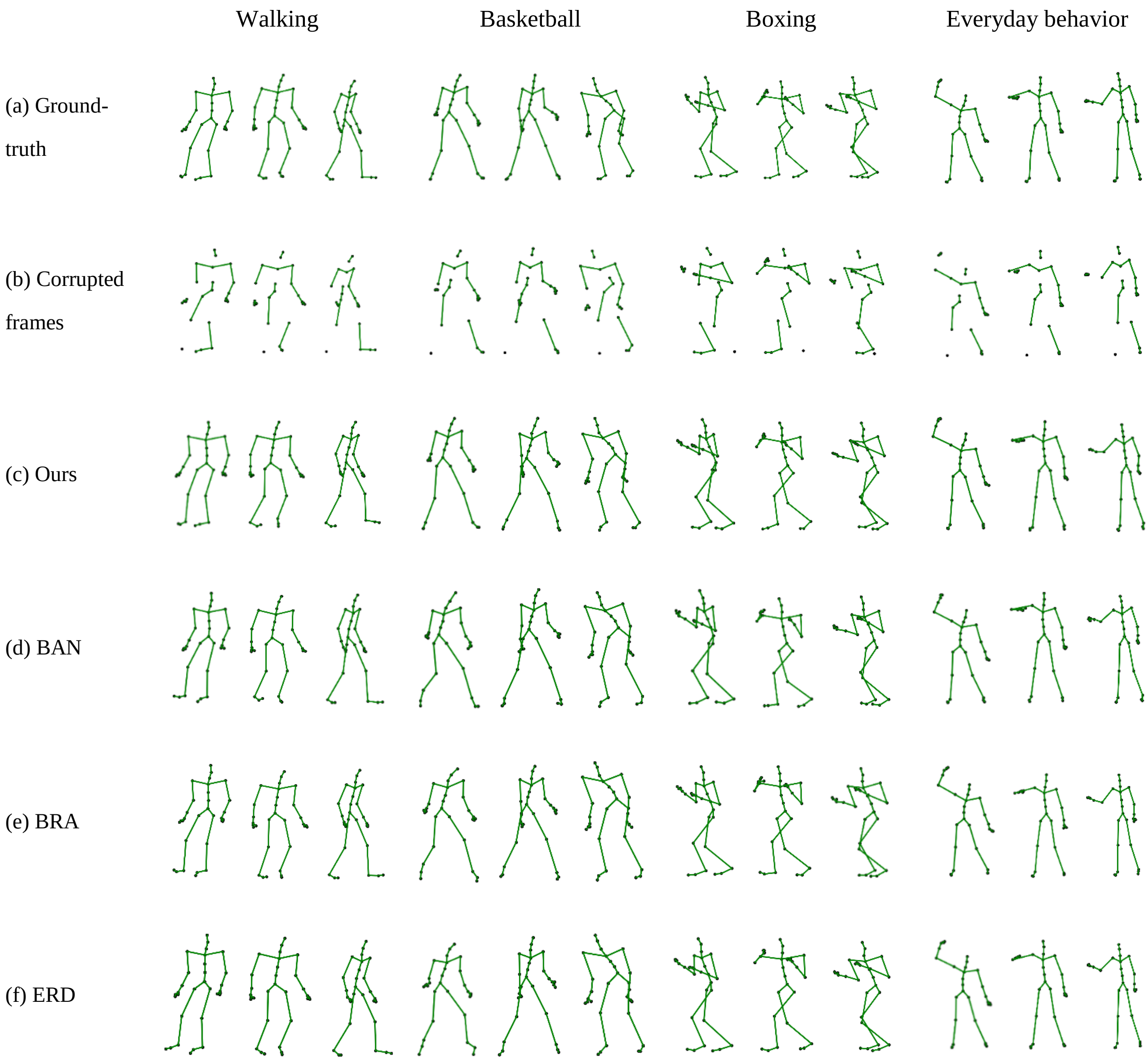


Figure 4. Denoising results of 30% missing data that were obtained via four approaches

Figure 4. give examples that are recovered via different method after randomly missing 30% markers. Figure 4(a) are the ground-truth frames. Figure 4(b) are the corrupted frames. Figure 4(c) The sequence after denoised via our method. Figure 4(d) to Figure 4(f) are the sequence after refinement via BAN, BRA and LSTM. As can be seen from the figure, for motion sequences with strong regularity such as walking, the recovered bones are very close to the original ones. The other sequences, although more or less a little unnatural, the reconstruction motion sequences by we proposed are also robust and outperform the reference method.

Repairing continuously missing markers information is a difficult and challenging task for MoCap data cleanup. To evaluate the robustness, we segmented the motion sequences at 120 frames, and then removed from the motion segments at six markers respectively, with missing gaps of 40 frames for mild loss, 80 frames for moderate loss and 120 frames for severe loss. Finally, the motion sequence recovered by the four approaches were used to evaluate the performance. The quantitative comparisons of RMSE are shown in Table 2.

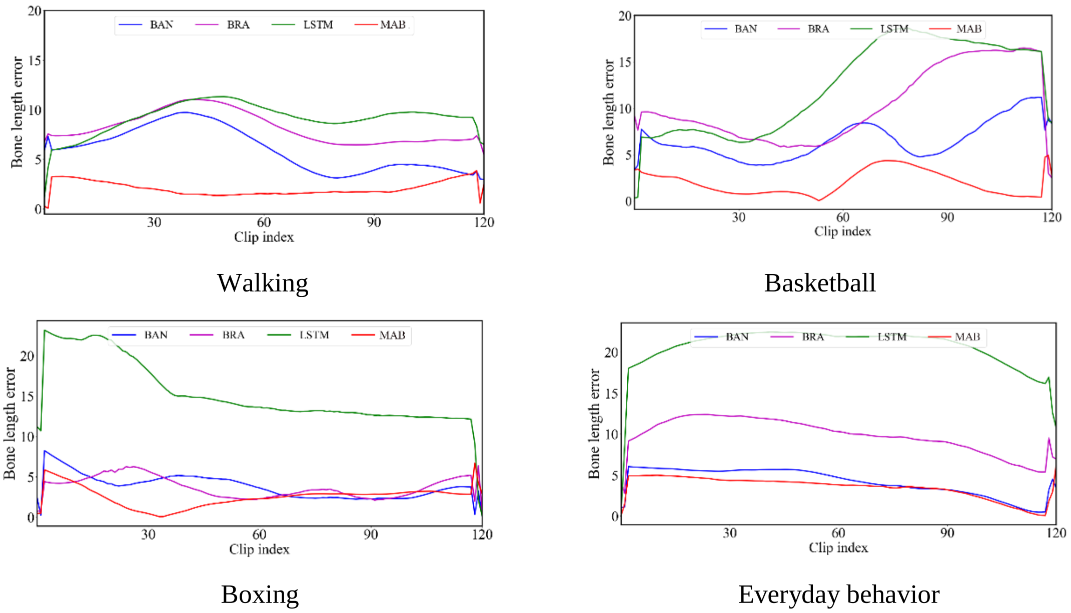
A comparative analysis of Table 2 shows that successive motion poses adjacent to the corrupt frames are lost when markers are missing for long periods of time, and the recurrent memory unit can only store short-term temporal correlations, so the recovery accuracy of each method steadily decreases with increasing gap time. However, our method is outperformed than the reference methods in most scenes. As the missing gap increases, even in the extreme condition where large-scale markers motion is severely impaired, the recovery performance of our method becomes more consistent. As shown in Figure 3, the RMSE achieved by our purposed method is relatively low when missing gap is 120 frames, and thus has better robustness to long gaps.

Table 2. Quantitative comparisons of RMSE of our method and the reference methods on four motions with different missing gap

Motion	Walking			Basketball			Boxing			Everyday behavior		
Missing gap	40	80	120	40	80	120	40	80	120	40	80	120
OURS	0.57	0.94	1.27	1.06	1.64	2.16	0.74	1.27	1.72	0.77	1.01	1.29
BAN	0.57	0.97	1.33	1.16	1.82	2.29	0.8	1.3	1.82	0.81	1.28	1.66
BRA	0.54	0.95	1.34	1.21	1.97	2.57	0.74	1.31	1.87	0.86	1.32	1.79
LSTM	0.61	1.04	1.4	1.18	1.90	2.45	0.77	1.37	1.98	0.84	1.27	1.7

Table 3. Quantitative comparisons of BLE of our method and the reference methods on four motions with different missing gap

Motion	Walking			Basketball			Boxing			Everyday behavior		
Missing gap	40	80	120	40	80	120	40	80	120	40	80	120
OURS	1.1	1.43	2.04	1.35	1.67	1.95	1.82	1.95	2.52	2.04	3.29	3.5
BAN	3.3	5.29	5.96	2.25	4.12	6.49	1.96	2.73	3.66	2.66	3.57	4.16
BRA	3.39	6.14	8.13	2.9	5.42	10.26	2.32	3.37	3.75	4.99	8.05	9.8
LSTM	3.45	6.29	9.13	3.38	6.76	12.3	6.98	11.35	15.1	7.59	14.37	20.5



Walking

Basketball

Boxing

Everyday behavior

Figure 5. Comparisons of BLE of our method and the reference methods on four motions with gap=120

In the process of motion reconstruction, if only the difference in the position of the markers recovery is considered, the output motion sequence is likely to be non-smooth, with some jitter, and it is difficult to produce the desired repair results using only the position loss, which needs to be judged with the help of the length of the bone. Table 3 give the bone length error for various methods with a random missing of six markers and missing gap of 40, 80, 120 frames, respectively. Figure 5. shows the bone length error curves for the four approaches at a gap value of 120 frames. The comparison of the experimental results reveals that the loss of bone length from our method is smaller in most scenarios. Thus, the proposed network architecture MAB not only obtain long-term dependencies, but also effectively maintains the natural, smooth motion.

5 Conclusion

This paper proposes an optimized neural network model MAB based on the attention mechanism, and constructs a loss function based on marker positions and bone length to improve efficiency of prediction, so that the model has a better repair effect even when large-scale markers are missing. However, the addition of the attention mechanism also brings a certain increase in computational time, which is a challenge for motion data capture systems with real-time high-performance requirements. Therefore, further exploration of this issue will be focused on in future research.

Acknowledgements

This work was supported by the Ministry of Education of Humanities and Social Sciences Project-- “Research on Motion Capture Technology in 3D Animation Creation” (Project No. 17YJC760124).

References

[1] Ø. Gløersen, P. Federolf, Predicting missing marker trajectories in human motion data using marker intercorrelations, *PLoS ONE*, Vol. 11, No. 3, Article No. e0152616, March, 2016.

[2] Y. Motai, S. K. Jha, D. Kruse, Human tracking from a mobile agent: optical flow and Kalman filter arbitration, *Signal Processing: Image Communication*, Vol. 27, No. 1, pp. 83-95, January, 2012.

[3] R. Y. Q. Lai, P. C. Yuen, K. K. W. Lee, Motion Capture Data Completion and Denoising by Singular Value Thresholding, *Eurographics 2011 (Short Papers)*, Llandudno, UK, 2011, pp. 45-48.

[4] D. Holden, J. Saito, T. Komura, T. Joyce, Learning motion manifolds with convolutional autoencoders, *SIGGRAPH Asia 2015 Technical Briefs*, Kobe, Japan, 2015, pp. 1-4.

[5] K. Fragkiadaki, S. Levine, P. Felsen, J. Malik, Recurrent network models for human dynamics, *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 4346-4354.

[6] S. Li, Y. Zhou, H. Zhu, W. Xie, Y. Zhao, X. Liu, Bidirectional recurrent autoencoder for 3D skeleton motion data refinement, *Computers & Graphics*, Vol. 81, pp. 92-103, June, 2019.

[7] A. P. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, June 2016, https://arxiv.org/abs/1606.01933.

[8] S. J. Howarth, J. P. Callaghan, Quantitative assessment of the accuracy for three interpolation techniques in kinematic analysis of human movement, *Computer Methods in Biomechanics and Biomedical Engineering*, Vol. 13, No. 6, pp. 847-855, December, 2010.

[9] H. J. Shin, J. Lee, S. Y. Shin, M. Gleicher, Computer puppetry: An importance-based approach, *ACM Transactions on Graphics (TOG)*, Vol. 20, No. 2, pp. 67-94, April, 2001.

[10] C. C. Hsieh, P. L. Kuo, An impulsive noise reduction agent for rigid body motion data using B-spline wavelets, *Expert Systems with Applications*, Vol. 34, No. 3, pp. 1733-1741, April, 2008.

[11] J. Xiao, Y. Feng, M. Ji, X. Yang, J. Zhang, Y. Zhuang, Sparse motion bases selection for human motion denoising, *Signal Processing*, Vol. 110, pp. 108-122, May, 2015.

[12] C. Yen, G. Chen, A Deep Learning-Based Person Search System for Real-World Camera Images, *Journal of Internet Technology*, Vol. 23, No. 4, pp. 839-851, July, 2022.

[13] S. Guo, Z. Wang, Y. Lou, X. Li, H. Lin, Detection Method of Photovoltaic Panel Defect Based on Improved Mask R-CNN, *Journal of Internet Technology*, Vol. 23, No. 2, pp. 397-406, March, 2022.

[14] U. Mall, G. R. Lai, S. Chaudhuri, P. Chaudhuri, A deep recurrent framework for cleaning motion capture data, December, 2017, https://arxiv.org/abs/1712.03380.

[15] T. Kucherenko, J. Beskow, H. Kjellström, A neural network approach to missing marker reconstruction in human motion capture, March, 2018, https://arxiv.org/abs/1803.02665.

[16] L. H. Son, A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, M. Abdel-Basset, Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network, *IEEE Access*, Vol. 7, pp. 23319-23328, February, 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems*, Long Beach, CA, USA, 2017, pp. 5998-6008.

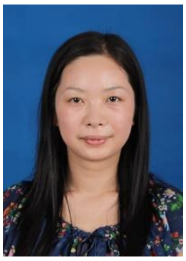
[18] F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, Se (3)-transformers: 3d roto-translation equivariant attention networks, *Advances in Neural Information Processing Systems* 33, Vancouver, BC, Canada, 2020, pp. 1970-1981.

[19] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, USA, 2019, pp. 3085-3094.

[20] Q. Cui, H. Sun, Y. Li, Y. Kong, Efficient human motion recovery using bidirectional attention network, *Neural Computing and Applications*, Vol. 32, No. 14, pp. 10127-10142, July, 2020.

[21] M. Li, Z. Miao, W. Xu, A CRNN-based attention-seq2seq model with fusion feature for automatic Labanotation generation, *Neurocomputing*, Vol. 454, pp. 430-440, September, 2021.

Biographies



Yongqiong Zhu is an associate professor in the School of Art, Wuhan Business University, Wuhan, China. She received Ph.D. degree of Engineering in Computer Science School from Wuhan University, Wuhan, China, majoring in Communication and Information System. Her research interesting areas include Motion capture, 3D reconstruction, Interaction design.

Fan Zhang received the Ph.D. degree in Computer Science School from Wuhan University. He is currently an associate professor at the school of Mathematics & Computer Science of Wuhan Polytechnic University, China. His research interests include Information system security, and Machine learning security.

Zhidong Xiao is currently a Principal Academic and Deputy Head of Department at National Centre for Computer Animation, Bournemouth University, UK. He received PhD degree in Computer Graphics and Computer Animation at National Centre for computer Animation, Bournemouth University. His research focuses on Motion Capture, Physics-based Simulation, Motion synthesis, Machine Learning and Virtual Reality.