Motion Capture Data Denoising Based on LSTNet Autoencoder

Yong-Qiong Zhu[1], Ye-Ming Cai[2], Fan Zhang[3*]
[1]School of Art, Wuhan Business University, China
[2]School of Computer Science, Wuhan University, China
[3]Mathematics and Computer Science School, Wuhan Polytechnic University, China
yongqiongzhu@163.com, 2018282110317@whu.edu.cn, whpuzf@whpu.edu.cn

Abstract

This paper proposes a novel deep learning-based optical motion capture denoising model encoder-LSTNet- decoder (ELD). ELD uses an autoencoder for manifold learning and decoder to remove jitter noise and missing noise effectively. It uses recurrent units in LSTNet to effectively obtain the spatial-temporal information of motion sequences, especially the periodic long-term and short-term features. In the denoising procedure, the kinetical characteristics of the motion are also considered so that the reconstructed deviation is smaller and can more accurately reflect the real action. We simulated ELD with the CMU database and compared it with the art-of-state methods. The experiment shows that ELD is a very effective denoising technique with lower reconstruction error, stronger robustness, and shorter running time.

Keywords: Deep learning, Motion capture, Manifold learning, Denoising

1 Introduction

Optical motion capture systems have been successfully applied to film art, game and animation, medical rehabilitation, motion analysis, military simulation training, and other fields.

Because the movement of luminescent markers is captured, the captured motion sequence may have some noise due to the error of calibration, the occlusion of the maker, and the low resolution of the sensor. There are two main types of noise. One is that the position of the marker is offset compared with the actual situation of the movement, and the jitter of a certain amplitude appears. The other is the loss of captured marker data. When the marker is missing for a long time, it isn't easy to recover the original motion accurately from the context.

Commercial motion capture systems (such as VICON) generally rely on interpolation methods for denoising, that is, linear interpolation or spline interpolation based on the movement trajectory of the marker. However, these interpolation methods often fail when the trajectory curvature of the marker changes greatly or disappears on a large scale, thus relying on manual adjustment by the animator. Aiming at this problem, researchers have proposed many methods. In [1], signal filtering is used to remove the impulse noise in the motion sequence. The motion data containing noise is decomposed to assign higher magnitude coefficients to the noise, and then the noise is removed by smoothing these high magnitude coefficients. In [2], the human motion sequence is represented by a low-rank matrix, so denoising is transformed into a problem of filling recovery using low-rank theory. [3] takes the data in the MOCAP database as prior data to discover and learn the potential spatial and temporal information in the motion sequences, and then makes denoising based on this information.

In recent years, with the wide application of deep learning algorithms in artificial intelligence [4], researchers have found that deep learning-based methods are very suitable for motion capture denoising, and many effective methods have been proposed [5-7]. [6] offers a fully connected neural network to learn past temporal and spatial information of motion sequences to predict missing data. [7] suggested using the recurrent neural network LSTM to deal with the noise of marker missing. In [8], the six-layer ResNets are used to remove the noise. However, fully connected neural networks and ResNet are unsuitable for processing temporal data and require much calculation. Although the RNN has a good advantage in processing temporal data, it cannot perform better motion reconstruction in large-scale marker missing [9].

In 2015, Holden et al. proposed to use an automatic encoder and decoder for manifold learning of human motion [10]. Since human motion data is usually represented by the Euler angle and position of joints, this kind of representation is very suitable for data processing. The manifold learning of human motion can be regarded as the prior probability distribution of human activity. He used the convolutional autoencoder to find an effective subspace in the motion sequence and applied it to denoise in the corrupted motion. Inspired by Holden, [11] proposed an encoder-rnn-decoder network (ERD), then [9] proposed the EBF model of encoder-blstm-decoder. However, the experiment shows that these proposed models are not general and robust, the reconstructed results are not natural, and the EBF costs much time [12]. In 2018, [13] proposed a neural network model LSTNet specially designed for long-term and short-term multivariable temporal prediction tasks. The advantage of LSTNet is that it can use the convolutional layer to obtain the short-term local characteristics of temporal data and use the recurrent layer to discover the long-term features. LSTNet is very suitable for noise prediction of human motion sequences because it also has long-term and short-term periodic characteristics.

Therefore, inspired by work [10, 13], this paper presents an autoencoder neural network architecture named ELD. It builds three layers fully connected neural network as an encoder to perform the manifold learning of human motion. Furthermore, it makes three layers fully-connection decoder to reverse projection to learn the movement characteristic and

* Corresponding Author: Fan Zhang; E-mail: whpuzf@whpu.edu.cn
DOI: 10.53106/160792642022012301002

dimension transformation. As a result, it greatly improves the refined accuracy. In addition, compared with LSTM, LSTNet effectively obtains the long-term periodic motion features and has a better effect on the non-periodic prediction, so the reconstructed motion sequence can better reflect the real action. Finally, the denoising speed is fast, which is suitable for real-time applications. The method's results are shown in Figure 1.

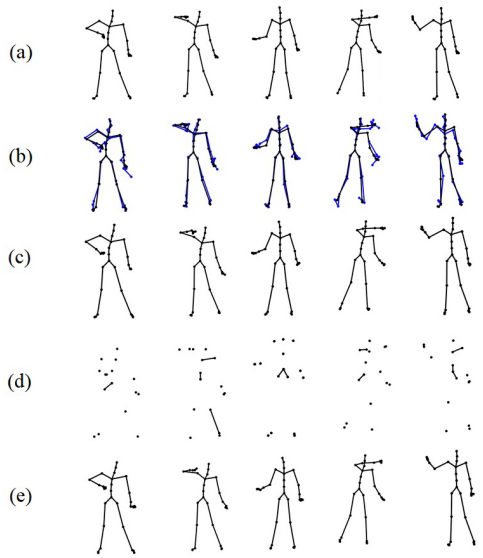


Figure 1. Examples of denoising via our method. (a) A ground-truth motion sequence. (b) Gaussian noise sequences use blue lines and the black lines is a corresponding ground-truth sequence. (c) The sequence obtained by refining gaussian noise using our method. (d) A corresponding motion sequence that is randomly removed 50% markers. (e) The sequence in (d) after denoised via our method.

The main contributions of this paper are as follows:

1. An effective neural network denoising architecture based on deep learning is proposed. The network model can effectively obtain the temporal and spatial information of the motion sequence, especially the periodic long-term and short-term characteristics. Furthermore, it uses a multi-layer fully connected layer as an encoder for manifold learning and the decoder reverse projection to remove jitter noise and missing noise.

2. The combined loss function of position and bone length is adopted to make the reconstructed motion sequence more stable and more accurate to reflect the real action.

3. The cost of this method is small, and it is more suitable for real-time applications.

The structure of this article is as follows. In the second part, the related work about denoising and model based on deep learning are investigated. The third part introduces denoising and the network model, loss function, and associated theoretical analysis. In the fourth part, experiments based on the CMU motion capture database are designed to verify the effectiveness of the proposed method, and a detailed investigation is carried out. Finally, in the fifth part, the conclusion of this paper is given.

2 Related works

For the problem of denoising MOCAP data, researchers have proposed many methods, mainly divided into three categories: filter-based method, matrix-based method, and data-driven method.

1. Filter-based method

In 2001, [14] introduced the Kalman filter method to map human motion to CG role in real-time for the first time. In 2009, [15] used the Kalman filter to extract the internal relationships of the motion sequences to fill in the missing data. According to their later work [16], if only linear dynamic systems and Kalman filtering are used without considering motion constraints, it will produce impractical results. So, they set position constraints to improve performance. In 2016, [17] projected the motion data into a low-dimensional space, then use Kalman filter and low-rank matrix to complete the refinement of missing markers.

The filter-based method is a signal filtering method, which is very fast and effective when the corrupt data is small-scale. However, these methods do not use the human body structure information implicit in MOCAP data -- temporal and spatial information. Therefore, the context information cannot be retained effectively, so it cannot deal with large-scale corrupt data. In addition, when the time interval of the missing marker is long, this method is usually invalid and requires manual intervention [18].

2. Matrix-based method

There are two main denoising methods based on the matrix: principal component analysis (PCA) and low-rank matrix.

Since the impact of random noise on the original data is relatively small and all components are independent of each other, PCA operation can retain the principal elements and eliminate the noise to reconstruct the original data. [19] proposed to use the low-dimensional local linear model characterized by main components to establish the training set of the motion sequence. For a new motion containing missing markers, pre-trained classifiers can be used to identify the most appropriate local linear model for each frame. The least-squares solution is used to recover the missing markers based on the available marker positions and the principal components of the related models. The advantages of this model are that it is very simple and fast, and few parameters need to be adjusted. The same year, [20] proposed the GKPCA algorithm, which filtered out part of samples based on a greedy algorithm. Compared with PCA, GKPCA can reduce the training set by filtering the feature space to minimize the computational amount and the cost of denoising.

The PCA-based method assumes that the training set can represent the motion data space, but it cannot fully sample the space. Moreover, PCA only considers the change of the data in the orthogonal direction, which leads to accuracy decrease after refinement.

Another denoising method using the matrix is based on the low-rank matrix theory. If a matrix represents the human motion sequence, the matrix is a low rank, which can be proved by calculating the eigenvalues of the matrix. Based on this discovery, [2] proposed to use low-rank matrix theory to solve the denoising problem of MOCAP data in 2011. When some items of the motion sequence matrix are missing, a singular value thresholding algorithm (SVT) is used to fill the matrix. This method proves the validity of using low-rank attributes and does not require a motion prior. In 2014, [21] also considered the motion data's low-rank structure and time stability, added the smoothness constraint and used the Augmented Lagrange Method (ALM) algorithm to accelerate

the matrix solution and improve the calculation performance. In 2016, [22] used sparse representation to retain statistical information and applied smoothness constraints to have kinetical information.

The advantages of these methods are that the denoising speed is very fast, but if an entire row or column in the low-rank matrix is lost, it is impossible to complete the matrix reconstruction. That is, it is impossible to reconstruct the severely damaged human motion data.

3. Data-driven approach

In 2011, [3] first proposed a data-driven MOCAP data denoising method. He took the motion data as prior data and established a globally effective spatial index structure, K-nearest-neighbor tree (KD tree). Through the retrieval of the KD tree, it can find the missing marker. This method can save all motion data in a prior database and can be extended well. In recent years, with the development of machine learning, the idea of training machine learning models for denoising by learning large-scale data has been applied to many problems in computer graphics, such as denoising for data recovery or painting tasks [23]. Many methods based on deep learning are proposed [24].

In 2017, [6] proposed the structural recursive neural network (S-RNN) based on a space-time diagram, which focuses on the interaction between environment and human and has good generalization, providing a new method for refining missing markers. In 2018, [7] proposed a fully connected network based on a sliding window. It uses the information in the sliding window to predict the coordinates of missing markers in the future window. But the longer the window, the harder the performance bottleneck becomes. In the same year, [25] proposed a method based on self-similarity analysis, the essence of which is to use the KNN algorithm to get the K-frames most similar to the frames to be repaired and calculate the position estimate by weight. However, if there is no corresponding relationship between the two frames, the KNN matrix not be established. Then, [8] uses ResNets to map corrupt markers to the corresponding joint transformation. However, the processing time is long and is unsuitable for real-time applications.

There is another kind of denoising method based on the neural network of the autoencoder. As the dimension of the human motion sequence model is very high, [10] find that manifold learning can reduce its dimension. It uses the automatic convolutional encoder to transfer the local relations of high-dimensional data to the low-dimensional hidden space, to obtain the low-dimensional representation of human motion. Then acquired the low-dimensional training features. The work also demonstrates that projection and reverse projection using encoder and decoder can be applied to refined corrupt motion data such as Gaussian noise or missing noise. However, only using the auto-convolutional encoder will result in a motion jitter [9]. In the same year, [11] proposed an ERD network based on [26] by adding nonlinear fully connected layers to the LSTM network as the encoder and decoder. According to [11], the motion capture data is typical of the spatial-temporal feature, and a combination of nonlinear encoder and recurrent units can result in better prediction. However, the experiment shows they are not robust, and the reconstructed motion is not natural [27]. In 2017, inspired by the ERD model, [9] proposed the EBF model, whose main improvement is to change the recurrent network into a bidirectional recurrent network and use four fully connected layers in the encoder design. [9] also constructed an EBD model similar to EBF for human skeleton reconstruction and then used the EBF model to denoising. However, it costs a lot of time to build a skeleton according to all frames. In 2019, [28] proposed BRA network architecture inspired by [10-11]. BRA is also based on the BLSTM network but differs from [9] in encoder design. In 2020, [29] used the same BLSTM network as [28]. The difference in [28-29] is that [28] increased the smoothness loss in the motion reconstruction. At the same time, [29] introduced the Attention mechanism [30] in the encoder, aiming to make the reconstructed motion more natural by imitating human habits. However, [28] could not effectively recover the large-scale missing marker for a long time, and the robustness was poor, while [29] had a poor recovery for the jitter noise. In addition, although the gradient disappearance problem has been alleviated to some extent in BLSTM, it is still difficult for LSTM to capture long-term features when the missing time is very long, resulting in low accuracy of noise detection.

In summary, using deep learning technology for motion capture denoising is an effective method. However, the currently proposed neural network architecture still has many shortages in the accuracy, robustness, and time cost of denoising. Therefore, inspired by [10, 13, 28], this paper proposes an automatic encoder denoising framework based on LSTNet. The framework can effectively capture long-term motion features, alleviate the "gradient disappearance" and "gradient explosion" problems, and the automatic encoder can perform effective manifold learning of motion sequences without human manual intervention.

3 Problem Solution

In this part, 3.1 introduces the pipeline of denoising and the network architecture is given in 3.2. Then, the corrupt data generation is in 3.3. Finally, the network training process is shown in 3.4.

3.1 Problem Description

A human motion capture data consists of a series of frames, each recording the three-dimensional coordinate positions of various joints in the human skeleton. Assume $X=\{x_1, x_2, \cdots, x_t, \cdots, x_n\}$ is the human motion capture matrix, where $1 \le t \le n$, x_t represents frame t, so $x \in R^{3d \times n}$ and d is the number of joints. Assume that the ground-truth sequence is x^G, the noise sequence is x^c, and the reconstructed motion sequence after denoising is $Y=f(x^c)$. Then the denoising problem is transformed into finding an optimization function that can minimize the difference between the refined motion data Y and the ground-truth sequence x^G.

$$\min_f \left\| Y - x^G \right\|_F^2 \tag{1}$$

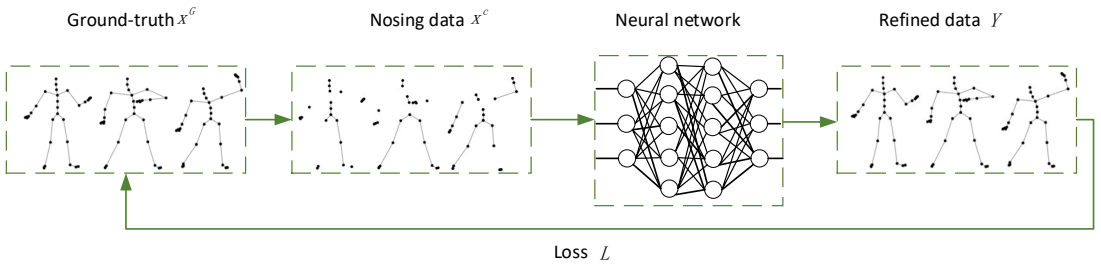


Figure 2. Denoising pipeline

The denoising pipeline is shown in Figure 2. Firstly, corrupt data X^c are obtained by adding noise to the original motion X^G. Then forward propagation in the neural network and the loss rate L is fed back to the neural network to update the weight and offset. Lastly,the corrupt data is tested to get the refined motion sequence Y.

3.2 Network Architecture

This paper proposes a neural network architecture ELD for automatic encoder based on LSTNet. The architecture is shown in Figure 3.

The automatic encoder uses a 3-layer fully connected neural network for manifold learning of MOCAP data. The first layer converts the output dimension of each frame to 128, the second layer converts the size to 256, and the third layer converts the dimension to 512, which is input to LSTNet.

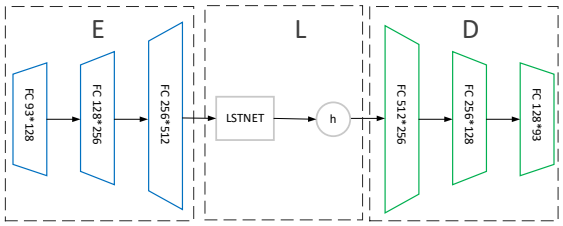


Figure 3. ELD architecture

LSTNet uses a convolutional and recurrent layer to obtain the motion sequence's periodic features. LSTNet introduces a recurrent-jump layer with a time-span and gets the regular part by extending the time-span of the data flow. For example, to predict the value at time t and use the most recent records, it is also based on the history of a time-span p to find the regularity. In this paper, the convolutional layer of LSTNet is implemented using one-dimensional convolution, with 48 filters, and the size of the convolution kernel is 2*93. The input data is output to the recurrent layer and recurrent -jump layer through the convolutional layer, each of which contains 512 hidden units.

The decoder design is composed of three fully- connected layers added after the hidden unit of the LSTNet. It integrates the eigenvalues between motion sequences learned by the neural network, and dimension transformation is carried out simultaneously to transform the eigenvectors into input dimensions. The first layer converts the output dimension of each frame to 256, the second layer converts the size to 128, and the third layer converts to 93. That is, the output vector is converted to the input dimension.

3.3 Noise generation

To simulate corrupt motion data, we need to add noise to the MOCAP data. The noise-adding algorithm is as Algorithm 1.

Algorithm 1. The algorithm is to add noise to the ground-truth X^G and return corrupt data X^c.

Function Corrupt(X^G, $SNR_{db} \in R$, $M \in R^{3d \times n}$)
  if (it has jitter noise)
    $noise = Guassian(SNR_{db})$
    $x^c = x^G + noise$
  else if(it has missing noise)
    $x^c = M \odot x^G$
  else return x^c
End

For jitter noise, we use Gaussian noise to simulate. Given the signal-to-noise ratio SNR_{db}, vector $noise$ is generated randomly based on the Gaussian noise. The higher the SNR_{db}, the less noise the signal contains.

For missing noise, a binary vector mask $M \in R^{3d \times n}$ is used to add noise to the ground-truth data X^G. M is a vector that the vector value is in (0,1). $\odot$ is a product of vectors. If a marker is missing, its value in the vector should be 0.

3.4 Training

Assume that x_t^c is the corrupt data at a time t .We input x_t^c to fed to the nerual network. When it pass through the autoencoder (AC) and get the output h_t^{ac}, h_t^{ac} will input to LSTNet and output h_t^{lsn} .Then it is given to the decoder (DC) to get h_t^{dc}, and finally, output prediction result Y is acquired by the activation function tanh. Assuming that the learning rate is θ and loss L is calculated in the forward propagation, the gradient descent algorithm Adam is used to update network parameters through backward propagation. The training process is shown in algorithm 2.

Algorithm 2. The ELD training

Function Train(x^c, $\theta \in R$)

$h_t^{ac} = AC(x_t^c)$

$h_t^{lst} = LstNet(h_t^{ac})$

$h_t^{dc} = DC(h_t^{lst})$

$Y = \tanh(h_t^{dc})$

compute L

$\theta = AmsGrad(\theta, \Delta L)$

End

In the design of loss function, it is commonly used the position of marker deviation to evaluate the loss. Define the position loss function between the predict marker position in Y and the ground-truth marker position in x^G as L_P:

$$L_P = \frac{1}{n}\left\|Y - X^G\right\|_2 \tag{2}$$

and $\left\|\square\right\|_2$ is $L2$ norm.

However, the experiment in [31] proves that only considering the position loss cannot deal with jitter noise effectively. If the loss function does not view the kinetical information of the motion, it will cause the action to be unsmooth. Therefore, it is difficult to get an ideal reconstructed motion sequence using only the loss function of the marker positions. Keeping the bone length of the reconstructed motion unchanged is a very important kinetical feature and is the key to denoising.

In this paper, we propose a loss function combining position and bone length. The loss of bone length was defined as L_B:

$$L_B = \sum_{i=1}^{n}\sum_{j=1}^{d}\left\|Y_{i,j} - x_{i,j}^G\right\|_2 \tag{3}$$

$Y_{i,j}$ is the length of the j bone in frame i of the refined motion sequence Y. $x_{i,j}^G$ is the corresponding bone length in the ground-truth motion. The loss function L is defined as:

$$L = \lambda_p L_P + \lambda_b L_B \tag{4}$$

$\lambda_p, \lambda_b \in (0,1)$. During network training, network loss L is minimized by adjusting the value of λ_p and λ_b.In our experiment λ_p is 0.98 and λ_b is 0.02.

4 Simulation

4.1 Dataset and preprocessing

The human motion capture data used in the experiment is from the CMU motion capture database. The data format is ASF/AMC, and the sampling frequency of ASF/AMC data is 120 frames/second. We randomly choose four types of motions, including two periodic motions (boxing, walk) and two non-periodic motions (everyday behavior and dancing). Every type of motion has been experimented with ten times, and the average value was recorded.

The computer used in this experiment has i9-10850K CPU and 32G physical memory, and the graphics card is Nvidia GeForce RTX3070. The programming language is Python3.7, and the deep learning framework is Tensorflow2.0.

To improve the convergence rate of the neural network, we need to map the MOCAP data to the range of [-1,1] firstly. The preprocessing is to convert the coordinate data of the joint of human motion into the coordinates with the root as the origin, find the average position of the motion sequence, and shrink the position of the joint to an interval centered on the average position. Finally, normalize the motion data to [-1,1] to get the data set. Then 60% of the data set is taken as the training set, 20% as the validation set, and 20% as the test set.

In the experiment, the batch size is 96, the epoch is 30, and the dropout is 0.5. Thus, in the recurrent layer, time-step is 5 and time-span p is 5 in the recurrent -jump layer.

4.2 Evaluation Indexes

The baseline methods are BRA [26] and ERD [10]. The main evaluation indexes include mean square error (RMSE) and bone length error (BLE).

$$RMSE = \frac{\left\|P_x - P_y\right\|_F}{\sqrt{d \times n}} \tag{5}$$

P_x represents the position of the ground-truth marker, P_y represents the position of the refined marker. $\|\cdot\|_F$ is F norm, d is data dimension and n is the number of frames in motion sequence. The smaller the RMSE value is, the smaller the difference between the reconstructed marker position and the original position , and the better the effect is.

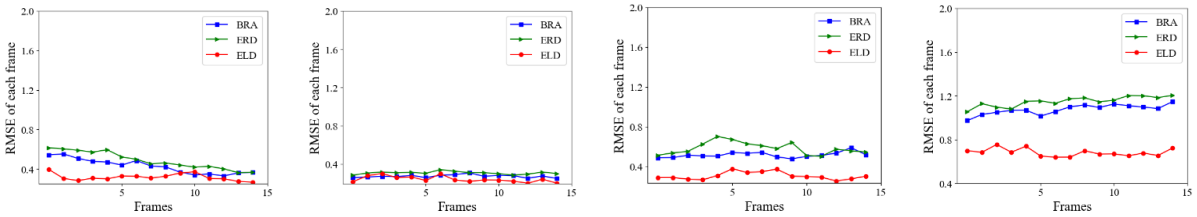
Bone length error :

$$BLE = \frac{\sum_{i=1}^{d}\left|B_i - \tilde{B}_i\right|}{n} \tag{6}$$

B_i is the bone length of each frame in the ground-truth motion , and $\tilde{B}_i$ represents the bone length of each frame in the corresponding refined motion. Bone length is calculated based on the position of the two joints connecting the bones.

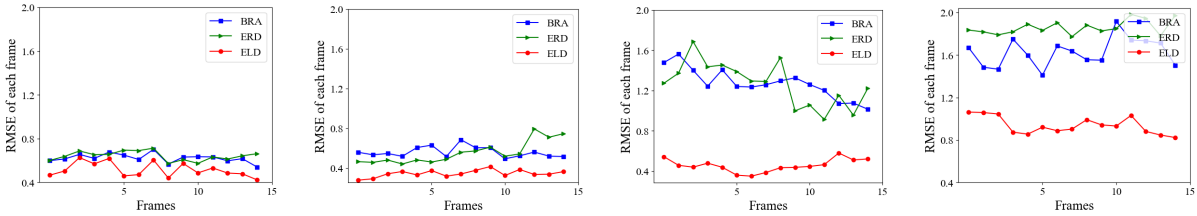
4.3 Experimental design and analysis

To verify the effectiveness of our model for the jitter noise, firstly , we experiment with the motion reconstruction comparison of different models when the Gaussian noise is 20dB, 10dB, and 1dB, respectively. That $SNR_{db}=20$ indicates that the jitter noise is low, and $SNR_{db}=1$ indicates that the jitter



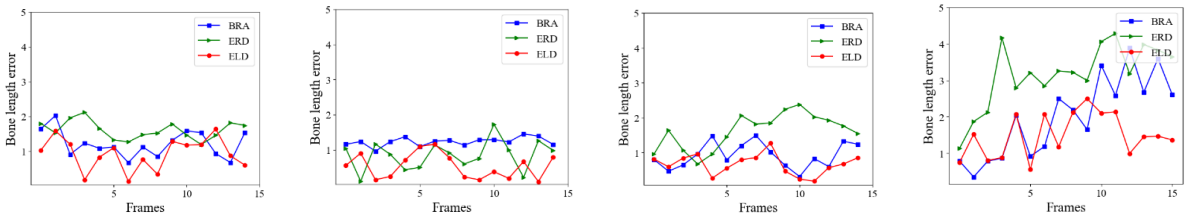
Boxing Walk Everyday behavior Dancing

Figure 4. Comparisons of RMSE between ELD and other two methods on four motions with SNR=20



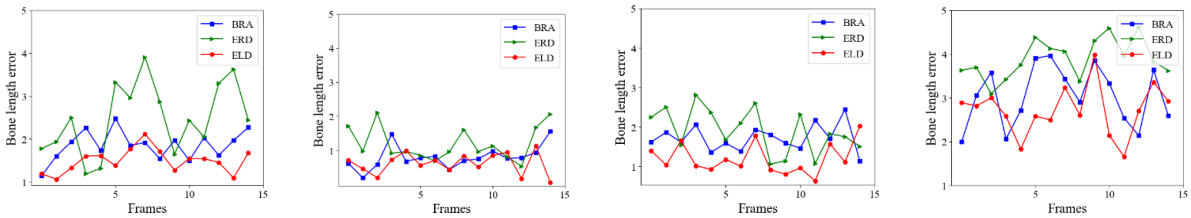
Boxing Walk Everyday behavior Dancing

Figure 5. Comparisons of RMSE between ELD and other two methods on four motions with SNR=1



Boxing Walk Everyday behavior Dancing

Figure 6. Comparisons of BLE between ELD and other two methods on four motions with SNR=20



Boxing Walk Everyday behavior Dancing

Figure 7. Comparisons of BLE between ELD and other two methods on four motions with SNR=1

is severe. Figure 4 shows the RMSE comparison results of four motions when $SNR_{db}=20$, and Figure 5 shows the comparison results of that $SNR_{db}=1$. The experiment also tested the bone-length comparison results , as shown in Figure 6 and Figure 7. Table 1 gives the comparison of RMSE values of each motion under different Gaussian noise amplitude, and Table 2 gives the the corresponding comparison value of bone length error.

From Figure 4 to Figure 7 and Table 1 and Table 2, we can conclude that: firstly, for each type of motion, our ELD method is better than the baseline models. Whether the noise amplitude is severe or slight , our RMSE and BLE are much lower than the baseline methods so that the reconstructed motion sequence will be closer to the ground-truth motion. Secondly, the RMSE and BLE of boxing and walk are relatively small for the aperiodic movement of everyday behavior and dancing because all models use recurrent units to capture the periodic motion features. But even for aperiodic activity, our method can greatly improve the prediction accuracy. Thirdly, the anti-jitter capability of BRA is better than ERD. This is because BRA design focuses on optimizing bone length, so the bone-length error and jitter level of ERD is higher than that of BRA.

In the actual capture process, jitter noise and missing noise often appear together. To verify the robustness of the model, we experiment with the prediction performance of long-term missing markers under the jitter noise SNR is 20. The motion sequence is segmented, and each segment is 120 frames. The missing marker is the knee joint, and the continuous missing frames Gap is set 30, 60 and 120 frames. Table 3 shows the RMSE values of different methods with different Gap, and Table 4 gives the corresponding BLE values. Figure 8 shows the frame-by-frame comparison of the RMSE values of each method with Gap=120, and Figure 9 shows the corresponding BLE values.

Table 1. RMSE between ELD and other two methods on four motions with different SNR

Motion	Boxing			Walk		
SNR	1	10	20	1	10	20
BRA	0.67	0.55	0.42	0.57	0.39	0.26
ERD	0.7	0.59	0.45	0.6	0.44	0.28
ELD	0.61	0.49	0.36	0.37	0.29	0.23
Motion	Everyday behavior			Dancing		
SNR	1	10	20	1	10	20
BRA	1.14	0.81	0.57	1.76	1.61	1.41
ERD	1.17	0.85	0.59	2.1	1.95	1.78
ELD	0.66	0.51	0.34	1.65	1.39	1.13

Table 2. BLE between ELD and other two methods on four motions with different SNR

Motion	Boxing			Walk		
SNR	1	10	20	1	10	20
BRA	1.74	1.41	1.21	1.0	0.87	0.79
ERD	1.95	1.77	1.6	1.87	1.51	1.15
ELD	1.61	1.32	1.09	0.84	0.69	0.53
Motion	Everyday behavior			Dancing		
SNR	1	10	20	1	10	20
BRA	1.76	1.59	1.41	3.64	3.51	3.35
ERD	2.1	1.92	1.78	4.31	4.22	4.0
ELD	1.65	1.38	1.13	3.14	2.65	2.13

Table 3. Comparisons of RMSE between ELD and other two methods on four motions with different numbers of continuous frames having missing joints

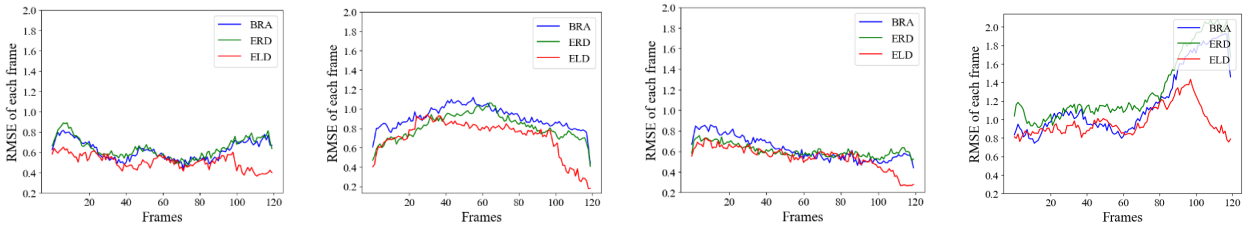
Motion	Boxing			Walk		
Missing Frames	30	60	120	30	60	120
BRA	0.44	0.47	0.62	0.35	0.47	0.9
ERD	0.47	0.49	0.64	0.34	0.45	0.81
ELD	0.36	0.42	0.54	0.29	0.42	0.72
Motion	Everyday behavior			Dancing		
Missing Frames	30	60	120	30	60	120
BRA	0.56	0.61	0.62	1.09	1.17	1.23
ERD	0.59	0.64	0.60	1.19	1.19	1.33
ELD	0.38	0.43	0.55	0.81	0.87	1.07

Table 4. Comparisons of BLE between ELD and other two methods on four motions with different numbers of continuous frames having missing joints

Motion	**Boxing**			**Walk**		
Missing Frames	30	60	120	30	60	120
BRA	1.54	1.65	2.19	1.75	1.79	2.05
ERD	1.19	1.61	1.7	1.2	1.28	1.35
ELD	0.92	1.15	1.47	0.79	1.03	1.09
Motion	Everyday behavior			Dancing		
Missing Frames	30	60	120	30	60	120
BRA	1.23	1.9	2.28	1.53	2.15	2.84
ERD	1.54	2.2	2.58	2.15	2.89	4.07
ELD	0.86	1.42	1.79	0.82	1.53	1.88

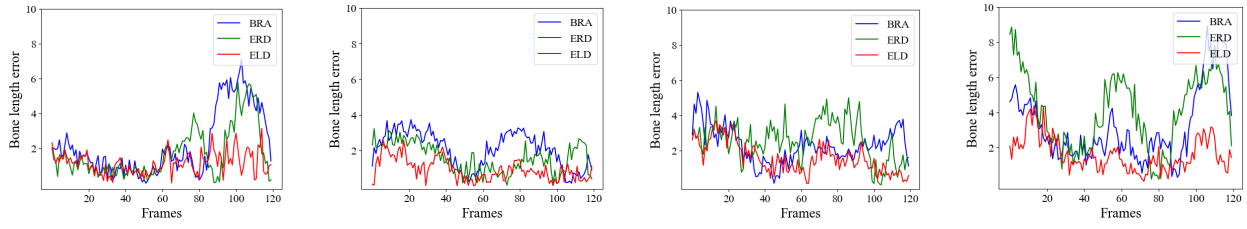
We can see from Table 3 and Table 4, and Figure 8 and Figure 9 that as the number of missing frames increases, the RMSE value and BLE value will continue to increase. However, compared with the reference methods, the error of the ELD is the smallest. The robustness is the best even when such a large-scale frame is missing. We can also find that the robustness of BRA is not better than that of ERD, indicating that BRA has a good effect against jitter noise, but it cannot effectively reconstruct large-scale missing noise.

Figure 10 shows the comparisons of the four types of motions under mixed noise. It can be seen that, due to the bone length constraint, the frame reconstructed by our ELD model is closer to the original frame than the reference methods. On the other hand, the recurrent layer and the recurrent jump layer in LSTNet can better capture long-term features, so the ELD's performance is better than the baselines.



Boxing Walk Everyday behavior Dancing

Figure 8. RMSE using different denoising methods with 120 continuous frames having missing joints and SNR=20



Boxing Walk Everyday behavior Dancing

Figure 9. BLE using different denoising methods with 120 continuous frames having missing joints and SNR=20

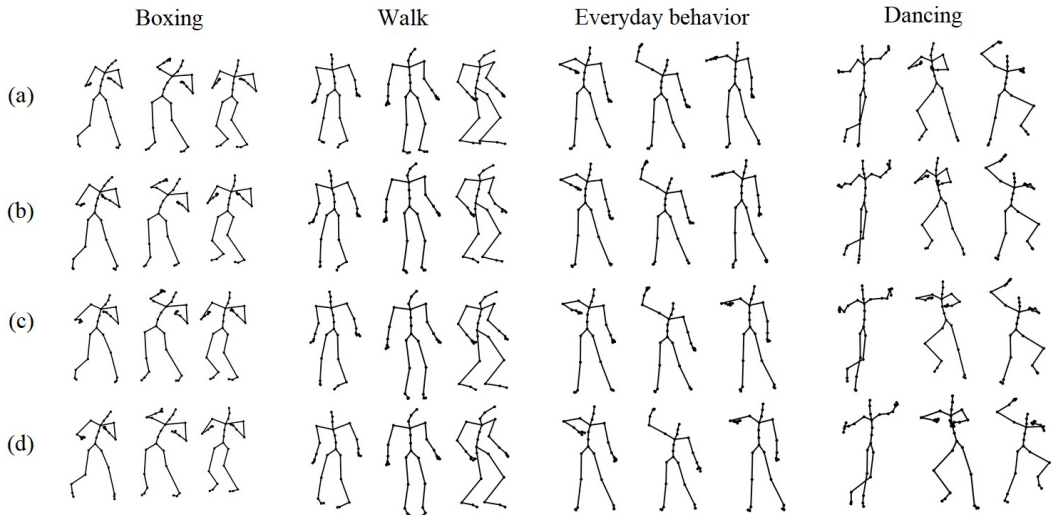


Figure 10. Reconstruction effects of four motions under Gap=120 and SNR=20
(a) Ground-truth data (b) ELD (c) BRA (d) ER

3. Processing time

In this experiment, we record the average time cost of training an epoch for each model. The results are shown in Table 5.

It can be seen from Table 5 that the training time of BRA is the longest, nearly twice as long as that of ERD, while the time cost of ELD is between them, close to BRA.

Therefore, this method can also be applied to real-time applications.

Table 5. The time of training in different models（Unit: second）

Motion	ELD	BRA	ERD
Boxing	1.654	1.774	0.894
Walk	1.582	1.733	0.851
Everyday behavior	1.833	1.944	1.092
Dancing	0.775	0.881	0.475

5 Conclusion

Because of the denoising problem in optical motion capture systems, researchers have proposed many neural network models to predict noise and refine it. However, the proposed model has poor performance. Therefore, this paper presents an ELD network model based on an automatic encoder, which effectively utilizes the temporal and spatial in context and kinetical information of motion to make the reconstructed movement closer to the real action and has good robustness. The performance of our method and the baselines are compared in the experiment. The results show that the ELD denoising model can get lower reconstruction error, stronger robustness, and shorter running time, which is a very effective denoising technology.

With the development of information technology [32-33] and artificial intelligence [34-35], future research will focus on denosing with fewer markers and making the reconstructed motion smoother.

Acknowledgements

This work was supported by the Ministry of Education of Humanities and Social Sciences Project (Project No. 17YJC760124).

References

[1] 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.

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

[3] J. Baumann, B. Krüger, A. Zinke, A. Weber, Data-Driven Completion of Motion Capture Data, *Workshop in Virtual Reality Interactions and Physical Simulation "Vriphys"*, Lyon, France, 2011, pp. 111-118.

[4] L. X. Liao, H. C. Chao, M. Y. Chen, Intelligently modeling, detecting, and scheduling elephant flows in software defined energy cloud: A survey, *Journal of Parallel and Distributed Computing*, Vol. 146, pp. 64-78, December, 2020.

[5] Y. S. Lin, C. F. Lai, C. L. Chuang, X. H. Ge, H. C. Chao, Collaborative Framework of Accelerating Reinforcement Learning Training with Supervised Learning Based on Edge Computing, *Journal of Internet Technology*, Vol. 22, No. 2 , pp. 229-238, March, 2021.

[6] J. Butepage, M. J. Black, D. Kragic, H. Kjellstrom, Deep representation learning for human motion prediction and classification, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 1591-1599.

[7] T. Kucherenko, J. Beskow, H. Kjellström, A neural network approach to missing marker reconstruction in human motion capture, *arXiv preprint arXiv:1803.02665*, March, 2018.

[8] D. Holden, Robust solving of optical motion capture data by denoising, *ACM Transactions on Graphics (TOG)*, Vol. 37, No. 4, pp. 165:1-165:12, August, 2018.

[9] U. Mall, G. R. Lai, S. Chaudhuri, P. Chaudhuri, A deep recurrent framework for cleaning motion capture data, *arXiv preprint arXiv:1712.03380*, December, 2017.

[10] 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.

[11] 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.

[12] T. B. Rodrigues, C. Ó. Catháin, D. Devine, K. Moran, N. E. O'Connor, N. Murray, An evaluation of a 3D multimodal marker-less motion analysis system, *Proceedings of the 10th ACM Multimedia Systems Conference*, Amherst, MA, USA, 2019, pp. 213-221.

[13] G. Lai, W. C. Chang, Y. Yang, H. Liu, Modeling long-and short-term temporal patterns with deep neural networks, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, Michigan, USA, 2018, pp. 95-104.

[14] 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.

[15] L. Li, J. McCann, N. S. Pollard, C. Faloutsos, Dynammo: Mining and summarization of coevolving sequences with missing values, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, 2009, pp. 507-516.

[16] L. Li, J. McCann, N. Pollard, C. Faloutsos, BoLeRO: a principled technique for including bone length constraints in motion capture occlusion filling, *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Madrid, Spain, 2010, pp. 179-188.

[17] M. Burke, J. Lasenby, Estimating missing marker positions using low dimensional Kalman smoothing, *Journal of biomechanics*, Vol. 49, No. 9, pp. 1854-1858, June, 2016.

[18] M. Perepichka, D. Holden, S. P. Mudur, T. Popa, Robust Marker Trajectory Repair for MOCAP using Kinematic Reference, *Motion, Interaction and Games*, Newcastle upon Tyne, UK, 2019, pp. 9:1-9:10.

[19] G. Liu, L. McMillan, Estimation of missing markers in human motion capture, *The Visual Computer*, Vol. 22, No. 9, pp. 721-728, September, 2006.

[20] T. Tangkuampien, D. Suter, Human motion de-noising via greedy kernel principal component analysis filtering, *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, China, 2006, Vol. 3, pp. 457-460.

[21] 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, *Information Sciences*, Vol. 277, pp. 777-793, September, 2014.

[22] G. Xia, H. Sun, G. Zhang, L. Feng, Human motion recovery jointly utilizing statistical and kinematic information, *Information Sciences*, Vol. 339, pp. 189-205, April, 2016.

[23] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, M. N. Do, Semantic image inpainting with deep generative models, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 6882-6890.

[24] C. Dai, X. Liu, J. Lai, P. Li, H. Chao, Human Behavior Deep Recognition Architecture for Smart City Applications in the 5G Environment, *IEEE Network*, Vol. 33, No. 5, pp. 206-211, September-October, 2019.

[25] A. Aristidou, D. Cohen-Or, J. K. Hodgins, A. Shamir, Self-similarity analysis for motion capture cleaning, *Computer graphics forum*, Vol. 37, No. 2, pp. 297-309, May, 2018.

[26] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Physica D: Nonlinear Phenomena*, Vol. 404, Article No. 132306, March, 2020.

[27] D. Holden, J. Saito, T. Komura, A deep learning framework for character motion synthesis and editing, *ACM Transactions on Graphics (TOG)*, Vol. 35, No. 4, pp. 1-11, July, 2016.

[28] 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.

[29] 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.

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

[31] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, G. Pons-Moll, Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time, *ACM Transactions on Graphics (TOG)*, Vol. 37, No. 6, pp. 1-15, November, 2018.

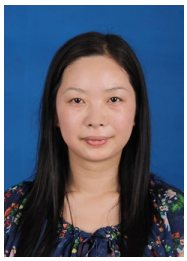
[32] S. Thombre, L. Patnaik, A. Tavildar, Delay and jitter sensitivity analysis with varying TCP fraction for multiplexed internet communications, International *Journal of Internet Protocol Technology*, Vol. 13, No. 2, pp. 61-77, February, 2020.

[33] D. Seo, B. Yoo, H. Ko, Information fusion of heterogeneous sensors for enriched personal healthcare activity logging, *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 27, No. 4, pp. 256-269, March, 2018.

[34] J. Wang, Y. Zou, P. Lei, R. S. Sherratt, L. Wang, Research on recurrent neural network based crack opening prediction of concrete dam, *Journal of Internet Technology*, Vol. 21, No. 4, pp. 1161-1169, July, 2020.

[35] B. Zeng, S. Yang, X. Yin, Robotic Relocalization Algorithm Assisted by Industrial Internet of Things and Artificial Intelligence, *Journal of Internet Technology*, Vol. 21, No. 5, pp. 1517-1530, September, 2020.

Biographies



Yong-Qiong 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.

Ye-Ming Cai received the master's degree in Data Analysis, University of Sheffield, UK, in 2020 and the master's degree in Computer Technology, Wuhan University in 2021. He has been working in Alibaba, China for algorithm develpoment.

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.