A Deep Bi-directional Attention Network for Human Motion Recovery

Qiongjie Cui, Huaijiang Sun, Yupeng Li and Yue Kong
Nanjing University of Science and Technology, Nanjing, China
{cuiqiongjie,sunhuaijiang}@njust.edu.cn, starli777@hotmail.com, codekong1028@163.com

Abstract

Human motion capture (mocap) data, recording the movement of markers attached to specific joints, has gradually become the most popular solution of animation production. However, the raw motion data are often corrupted due to joint occlusion, marker shedding and the lack of equipment precision, which severely limits the performance in real-world applications. Since human motion is essentially a sequential data, the latest methods resort to variants of long short-time memory network (LSTM) to solve related problems, but most of them tend to obtain visually unreasonable results. This is mainly because these methods hardly capture long-term dependencies and cannot explicitly utilize relevant context, especially in long sequences. To address these issues, we propose a deep bi-directional attention network (BAN) which can not only capture the long-term dependencies but also adaptively extract relevant information at each time step. Moreover, the proposed model, embedded attention mechanism in the bi-directional LSTM (BLSTM) structure at the encoding and decoding stages, can decide where to borrow information and use it to recover corrupted frame effectively. Extensive experiments on CMU database demonstrate that the proposed model consistently outperforms other state-of-the-art methods in terms of recovery accuracy and visualization.

1 Introduction

Human motion capture has gradually become the most popular motion storage technology in the industry, attracting a large number of scholars’ interest in research [Zhou et al., 2018; Bütepage et al., 2017; Mall et al., 2017]. It can be used in virtual reality, special effects movies, electronic games, and other related fields [Lu et al., 2018]. However, the raw mocap data may fail in completely recording the movement of all joints (including missing joint) due to inevitable reasons, such as marker falling off or joint occlusion. This inaccuracy and incompleteness of the captured data are often encountered even by professional motion capture equipment [Cui et al., 2019]. Further, corrupted motion sequences usually reveal a complex pattern in following aspects. First, the distribution of missing joints is unknown and arbitrary. Second, if the missing trajectory is too long, the information that can be used to repair the damaged motion will be insufficient. Third, the recovery accuracy will decrease rapidly in the case of large-scale movement (e.g., dancing, boxing). These factors present a major challenge for recovering the missing joints effectively.

Recently, some researchers have attempted to model human motion using deep neural networks [Mall et al., 2017; Holden, 2018]. They present various structures to solve related problems of recovering missing joints, which adequately analyze and utilize the spatio-temporal correlation of human motion [Gui et al., 2018]. Especially, the BLSTM-
based recurrent autoencoder [Mall et al., 2017] pave a golden path for modeling human motion. Although these models have made tangible progress, the performance may degrade rapidly over a long motion sequence because recurrent networks hardly capture the long-term temporal dependency and overcome the error accumulation problem. Besides, different motion frames should contribute unequally to the network while the previous models cannot consciously treat the context differently.

To address these aforementioned issues, we propose a deep bi-directional attention network (BAN) for motion recovery which leverages the attention mechanism and bi-directional long short-time memory network (BLSTM). Our inspiration comes from the recent theories of human attention which posit that human behavior can be efficiently modeled by the attention mechanism [Bahdanau et al., 2014; Zhou et al., 2016; Yang et al., 2016]. Specifically, the structure of our model consists of two components, encoder, and decoder, in which the attention mechanism is embedded to efficiently capture long-term temporal dependencies. In contrast to traditional attention, the proposed method adaptively calculates the relevant inputs of the forward and backward directions of BLSTM at the current time according to the correlation between the previous hidden state in both directions and all inputs. The long-term temporal dependencies are learned from chronologically arranged data and also from the reverse-chronological ordered data, which takes into account both forward and backward dependencies simultaneously. For human motion recovery, our BAN network explicitly selects the relevant context and selectively introduces the information from specific positions of the motion sequence to repair the damaged motion frame.

The specific contributions of this paper are summarized as follows: 1) We propose a novel bi-directional recurrent autoencoder for human motion recovery using attention mechanism. To our best knowledge, this is the first research attempt to exploit attention mechanism of BLSTM structure for human motion recovery. 2) We introduce the attention mechanism to efficiently capture long-term dependency and focus on the most important semantic information. 3) The experimental results demonstrate that the BAN achieves superior recovery accuracy and higher-quality visual results even for long-term motion sequences.

2 Related Work

Human motion recovery. Because of the inherent properties and structural constraints of human motion, the repair of missing joints cannot be simply regarded as data filling [Xia et al., 2018]. Many researchers have developed various methods to solve the problem of human motion recovery based upon the statistical properties (i.e., sparsity) of human motion [Lai et al., 2011]. Xiao et al. consider motion recovery from the perspective of sparse representation and propose a novel method named $l_1$-sparse representation (SR-L1) of missing markers prediction [Xiao et al., 2011]. Low-rank matrix completion [Tan et al., 2013; Hu et al., 2018], which usually seeks to find the lowest rank matrix for observed data, has been widely used for motion recovery. [Lai et al., 2011] first propose that damaged human motion can be efficiently recovered based on the low-rank prior, in which they use the singular value threshold method to solve the rank minimization problem. [Tan et al., 2013] suggest that mocap data based on trajectory representation can be used instead of frame representation, and the rank of this representation can be reduced because the lower rank is more suitable for the low-rank model. Nevertheless, these prior-based methods tend to yield unreasonable results for severely corrupted motion sequences. Because if the missing ratio is too large or the missing time is too long, the statistical property of low rank will no longer be satisfied.

Deep learning for human motion. Human motion is essentially a sequential data, which is naturally suitable for the sequential model in deep learning [Ruiz et al., 2018; Holden et al., 2017; Martinez et al., 2017]. Holden et al. develop various networks for human motion denoising and editing [Holden, 2018], but these structures abandon the temporal aspect of motion data. Alternatively, [Mall et al., 2017] propose a deep bi-directional recurrent network to clean up incomplete motion data wherein they use fully connected network to capture the joint correlation and temporal consistency of the human skeleton. [Fragkiadaki et al., 2015] present a recurrent autoencoder structure named Encoder-Recurrent-Decoder (ERD) to predict human body pose, in which they use LSTM [Hochreiter and Schmidhuber, 1997; Rumelhart et al., 1986] layer to learn temporal-spatial correlation of motion sequence. These structures achieve excellent results only in short-term sequences and cannot be directly used for recovering missing joints. For motion recovery, [Kucherenko and Kjellström, 2018] use a standard LSTM structure to recover motion sequence with missing markers in a short period of time.

Attention modeling. The seq2seq networks have produced stellar results, but one of the most challenging problems is the performance decline rapidly with the increase of sequence length [Bahdanau et al., 2014; Zhou et al., 2016; Yang et al., 2016]. To solve this problem, Bahdanau et al. adaptively select the relevant partially hidden state into the decoder at each time step using the attention mechanism [Bahdanau et al., 2014]. Yang et al., propose a hierarchical attention network for text classification using stacked recurrent layers, with each layer utilizing attention mechanism [Yang et al., 2016]. You et al., build an attention variant to learn to selectively tend to semantic concept proposals and integrate them into the recurrent neural network [You et al., 2016], which has achieved great success in the image caption. More recently, a dual-stage attention mechanism is proposed by [Qin et al., 2017] for time series prediction. In the first stage, the attention mechanism is equipped on a standard LSTM encoder to select the relevant inputs, while in the second stage the feature representation is also adaptively selected for decoding.

3 Methodology

3.1 Problem Formulation and Notation

In our work, a mocap matrix consists of a sequence of frames (poses), where each frame records 3D position of every joint and we formulate mocap data as $X =$
\( \{ x_1, x_2, \ldots, x_t, \ldots, x_n \} \in \mathbb{R}^{3d \times n} \) with \( 1 \leq t \leq n \). A frame at time step \( t \) is expressed as \( x_t \in \mathbb{R}^d \), where \( d \) is the number of markers in a human skeleton. We plan to solve the problem of motion recovery from the corrupted observation with missing markers. Assuming that \( X \in \mathbb{R}^{3d \times n} \) is the underlying complete mocap data, and let \( X_{\text{cor}} \in \mathbb{R}^{3d \times n} \) denotes a corrupted motion sequence. The values of the missing markers are recorded in the mocap data as 0 by the binary mask \( M \in \mathbb{R}^{3d \times n} \), i.e., \( X_{\text{cor}} = X \odot M \), where the symbol \( \odot \) denotes element-wise product. Then we transform the motion recovery task into optimizing a function \( g \) and \( f \) to minimize the difference between the recovered motion \( f(g(X_{\text{cor}})) \) and the complete motion sequence \( X \):

\[
\min_{f,g} ||X - f(g(X_{\text{cor}}))||. \quad (1)
\]

Then we use the autoencoder to fit the function \( f \) and \( g \). The encoder \( Y = g(X_{\text{cor}}) \) maps the observation \( X_{\text{cor}} \) into a low-dimensional representation, and then the decoder \( X_{\text{rec}} = f(Y) \) maps back into the input manifold to reconstruct the original signal.

### 3.2 Encoder with Bi-directional Attention

Our inspiration comes from human visual attention and BLSTM [Zhou et al., 2016; Qin et al., 2017; Bahdanau et al., 2014]. When humans receive a signal, they selectively receive stimulation as input in early stage. BLSTM, modeling sequential data from both forward and backward directions, has achieved superior performance. For motion encoder, not all frames contribute equally to the representation of BLSTM in the forward and backward directions. Therefore, we consider introducing attention mechanism into BLSTM to select the relevant input from two directions adaptively. The overall architecture of BAN is illustrated in Figure 2.

Assuming that the input sequence of the encoder is \( X = [x_1, x_2, x_t, \ldots, x_T] \), the bi-directional attention embedded in BLSTM includes forward and backward directions which can be built through a multi-layer perceptron. The calculation for the forward direction of BLSTM is formulated as:

\[
\begin{align*}
    e_{tf}^i &= v_{ef}^T \tanh(W_{ef}^T \tilde{h}_{t-1} + U_{ef} x_t), \\
    \alpha_{tf}^i &= \frac{\exp(e_{tf}^i)}{\sum_{i=1}^{T} \exp(e_{tf}^i)}, \\
    \tilde{x}_{tf}^i &= \sum_{t=1}^{T} \alpha_{tf}^i x_t,
\end{align*}
\]

where \( W_{ef} \in \mathbb{R}^{T \times 2m} \) and \( U_{ef} \in \mathbb{R}^{T \times 3d} \) are the weight matrix, and \( v_{ef} \) is a parameter to learn. \( \alpha_{tf}^i \) is the attention weight vector, which determines the importance of all inputs at time step \( t \). \( \tilde{h}_{t-1} \) and \( \tilde{x}_{t-1} \) are the hidden state and the cell state of forward BLSTM, respectively. With this process, we can extract the relevant input \( \tilde{x}_{tf}^i \) as the input of forward LSTM at each time step \( t \). Then, we can get \( \tilde{h}_{t} \) by newly weighted \( \tilde{x}_{tf} \), \( \tilde{h}_{t} = LSTM(\tilde{x}_{tf}) \). Similarly, the hidden state of backward LSTM can be calculated via \( \tilde{h}_{t} = LSTM(\tilde{x}_{ib}) \).

Finally, we obtain the hidden state at time step \( t \) by concatenating the forward hidden state \( \tilde{h}_{t} \) and backward hidden state \( \tilde{h}_{t} \), i.e.,

\[
h_t = \begin{bmatrix} \tilde{h}_{t} & \tilde{h}_{t} \end{bmatrix}.
\]

When we use the bi-directional attention mechanism to process the motion sequence, the encoder will adaptively select input frame through the importance of each frame for each direction of the bi-directional LSTM, instead of treating all frames equally.

### 3.3 Decoder with Bi-directional Attention

After all the frames are encoded, we will obtain a representation \( h \) of the corrupted motion. Then, the decoder maps the learned representation back into a recovered human motion.
We propose an adaptive input using bi-directional attention mechanism for correlation coding so that the decoder can pay attention to the most useful context in both directions along time step. As shown in Figure 2, the attention weight of each time step $t$ is calculated by the hidden state and the cell state of the previous time with:

$$
d^i_{tf} = v^\top_{df} \tanh(W_{df}[\tilde{c}^i_{t-1}; \tilde{c}^t_{t-1}] + U_{df} h^i_t),
$$

$$
\beta^i_{tf} = \frac{\exp(d^i_{tf})}{\sum_{i=1}^T \exp(d^i_{tf})},
$$

$$
\tilde{h}^i_{tf} = \sum_{i=1}^T \beta^i_{tf} h^i_t,
$$

where the $\tilde{c}^t$ and the $\tilde{c}^t$ denote the hidden state and cell state in forward LSTM of the decoder. $U_{df} \in \mathbb{R}^{T \times 2m}$ and $W_{df} \in \mathbb{R}^{T \times 2m}$ is the learnable weight matrix. $v^\top_{df}$ is a parameter vector that needs to be learn. We use the following formula to simply express this forward and backward process, i.e., $\tilde{c}^t = LSTM(\tilde{h}^i_{tf}), \tilde{c}^t = LSTM(\tilde{h}^i_{td}).$ Then, the hidden state of the decoder can be determined by concatenating $\tilde{c}^t$ and $\tilde{c}^t$, i.e.,

$$
c^t = [\tilde{c}^t, \tilde{c}^t].
$$

Finally, given the corrupted motion $X^{cor} = X \odot M$, the recovered motion sequence is obtained by feeding the semantic context $c_t$ into the time-distributed fully connected network. During training, our BAN takes $X^{cor}$ and $M$ as inputs, and then outputs the restored motion $X^{rec}$ at the same size as the input motion. Finally, we use the following formula to get the recovered motion:

$$
\bar{X} = M \odot X^{cor} + (1 - M) \odot X^{rec}.
$$

In particular, $\bar{X}$ is the weighted sum of $X^{cor}$ and $X^{rec}$. Notice that only the missing joint is reconstructed and the other parts are equal to the input.

3.4 Optimization

Let $X$ be the original motion sequence, $X^{cor}$ be corrupted motion, $M$ be the mask matrix, $X^{rec}$ be the recovered motion. We use two main losses to train our network. **Reconstruction loss**, that encourage the generator to preserve the information form the visible part of the sequence:

$$
\mathcal{L}_{rec} = \| M \odot X^{cor} + (1 - M) \odot X^{rec} - X \|_2.
$$

**Bone length loss**, which enforce constant bone length of the whole generated sequence:

$$
\mathcal{L}_{bone} = \sum_{i=1}^n \sum_{j=1}^d \| l^{rec}_{i,j} - l_{i,j} \|_2,
$$

where the $l_{i,j}$ denotes the $j$-th bone length of $i$-th frame of complete sequence, $l^{rec}_{i,j}$ is the corresponding segment of the recovered motion. The joint loss function is then formulated as:

$$
\mathcal{L}_{joint} = \lambda_{rec} \mathcal{L}_{rec} + \lambda_{bone} \mathcal{L}_{bone},
$$

where the $\lambda_{rec} = 0.95$ and $\lambda_{bone} = 0.05$ are the trade-off hyperparameters to fine-tune the importance of each loss term. They are determined by 10-fold cross validation.

4 Experiments

4.1 Dataset and Preprocessing

In this paper, we use CMU mocap database with 31 joint markers for the human body. Therefore, each frame can be represented as $X_t \in \mathbb{R}^{3 \times 31}$. We adopt the following methods to preprocess the mocap data.

(a) **Uniform height.** Before training, we scale all mocap data to achieve a uniform height. According to previous work [Holden, 2018], an appropriate scaling factor can be calculated by the average of all the bones of the actor.

![Figure 3: Qualitative results and visual comparisons with competitive methods in four different types of motion sequences.](image)

(a) Running  (b) Dancing  (c) Basketball  (d) Boxing

Figure 3: Qualitative results and visual comparisons with competitive methods in four different types of motion sequences. In each sub-figures, green segment are recovered parts and the unreasonable bone is circled with orange.
Figure 4: Quantitative comparisons of RMSE using different methods with 40% of continuous frames having missing joints

(b) Local reference system. We transform all the poses into the coordinate system with its root as the origin and use the y-axis in the world coordinate system as our y-axis. The x-axis is the horizontal direction from the left shoulder joint to the right shoulder joint. Then, the z-axis is produced by calculating the cross product between x-axis and y-axis.

(c) Normalization. We normalize the mocap data into the range [-1,1] by subtracting mean pose over the whole dataset and dividing into the absolute maximal value in the dataset.

The proposed model and other competitive methods are evaluated over the same configuration. During the training and testing, we randomly remove a certain number of active joints (10%, 20%, 30%, 40%), which is consistent with the situation of missing joints randomly. To simulate continuous missing of joint, we use gaps with a length of 60 frames where the total length of these gaps is 80% of the sequence, and randomly insert them into the mocap matrix. Such processing makes the position of the missing joint in the sequence random, while the length of the missing joint is 60 frames at the minimum and 80% of the sequence at the maximum, and concentrated at 40%, thus ensuring the randomness of the missing position and the continuous missing length simultaneously.

4.2 Implementation and Baselines

Our network uses BLSTM as decoder and encoder where each LSTM has 512 hidden units. The BAN model is trained using Adam [Kingma and Ba, 2014] with a learning rate of 0.001, and a more efficient mini-batch size 128 is applied to optimize the network. In our work, we use dropout [Srivastava et al., 2014] as the regularization method on the LSTM layer and the penultimate layer. With the dropout rate setting to 0.4, the model has better generalization performance. Note that the weights in our model are initialized randomly. The code will be available on the page: http://mocap.ai.

To better verify the performance of our model, we choose various methods for comparison, i.e., $l_1$ sparse representation (SR-L1) [Xiao et al., 2011], Low-rank matrix completion (LRMC) [Tan et al., 2013], long short-time memory network(LSTM) and window-based fully connected neural network (FCNN)[Kucherenko and Kjellström, 2018]. The hyperparameters are set to be consistent with those mentioned in their papers. Following the previous literature, the root means squared error (RMSE) measurement is adapted to quantify the recovered results:

$$\text{RMSE}(X_i, \tilde{X}_i) = \frac{1}{n_p} \left\| M_i \odot (X_i - \tilde{X}_i) \right\|^2_F,$$

where $X_i$ is the complete pose, $\tilde{X}_i$ is the recovered pose, and $n_p$ is the total number of imperfect entries (i.e., missing entries). $M_i$ is the degradation operator which removes all of the non-missing joints. Also, the bone-length error is an important criterion to determine whether the recovered motion sequence is visually reasonable and natural, i.e.,

$$\text{BLE}(X_i, \tilde{X}_i) = \frac{1}{n_p} \left| L_i - \tilde{L}_i \right|,$$

where the $L_i$ and $\tilde{L}_i$ are the sum of all bone length of $i$-th frame of recovered and corrupted motion, respectively.

4.3 Comparisons of Recovery Results

We first animate the comparison between our method and the competitive methods when the various type of motion sequences randomly lose 40% joints, such as, running, basketball, dancing, and boxing. In this case, both the missing markers and missing time are random. For each type of motion, we also pick out the character animation by visual recovery results corresponding to the five different moments of our method and the baseline methods. In Figure 3, the green segment is the recovered joint, and the blue part is the original joint position. In each sub-figure, the first row is the original motion, the second row is the corrupted motion, and the other rows are the recovery results of different methods. It is noteworthy that the recovered frame by the BAN is very similar to the original frame in most motion types, and the results obtained by our model are still robust even in the case of large-scale movements (e.g., boxing). However, the motion recovered by other methods is more or less unnatural and lacks visibility, which may lead to a failure of the recovery.

In the practical motion capture process, it is frequent for a certain joint to be lost continuously over a period, and the gap caused by this situation is difficult to handle. To simulate this situation, we continuously remove several active joints for each sequence. In Figure 4, we find that our method is more accurate than other methods regarding recovery error. Besides, as the missing time increases, the recovery accuracy of our method is more stable because bi-directional attention makes better use of the time context which allows the BAN to consciously determine where to borrow relevant information and use it reasonably. However, other competitive methods are susceptible to the number of missing frames.

As shown in Figure 3, there are many unreasonable bone lengths in the recovery results of the baseline methods under strenuous exercise, i.e., dancing, boxing. Specifically, we mark out unreasonable bone fragments in the animation with small circle in Figure 3. The motion recovered by our method is more accurate, and the bone length is more natural, which
makes the recovery result more in line with the real visual pose. We also measure the average bone-length error (BLE) for each motion sequence with 40% number continue frame having missing joints. From the quantitative comparison in Table 2, we can see that the bone-length error recovered by BAN is very small, and such a small error is difficult to observe in the actual animation. This means that under such a bi-directional attention mechanism, our model can find the most relevant context from the motion sequence to repair the damaged frame.

Due to the uncertainty and diversity of human motion, long-term human motion recovery is a challenging problem. To examine the limits of our method, we conduct a set of stress experiments to test recovery error in longer missing time, though occurring less frequently. We select 8 motion sequences. This indicates that the proposed method significantly improves the performance of human motion recovery concerning accuracy and visualization results, even in the case of long sequences or different missing distributions. However, there are still two defects that cannot be ignored: High time consumption, because the LSTM encoding and attention weight computation are non-parallel; Performance degradation for handling high missing ratio (> 80%). Fortunately, the cases of high missing ratio rarely occur in real-world applications. In the future work, we plan to use the idea of the generative model to further expand the scope of application of the proposed model and consider applying it to other tasks of human motion.

<table>
<thead>
<tr>
<th>Motion</th>
<th>Running</th>
<th>Climbing</th>
<th>Basketball</th>
<th>Boxing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing time</td>
<td>500</td>
<td>1000</td>
<td>1500</td>
<td>2000</td>
</tr>
<tr>
<td>SR-L1</td>
<td>2.73</td>
<td>2.98</td>
<td>3.32</td>
<td>3.35</td>
</tr>
<tr>
<td>LRMC</td>
<td>2.17</td>
<td>2.89</td>
<td>2.69</td>
<td>3.33</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.68</td>
<td>0.85</td>
<td>0.92</td>
<td>0.78</td>
</tr>
<tr>
<td>FCNN</td>
<td>0.98</td>
<td>1.11</td>
<td>1.15</td>
<td>1.09</td>
</tr>
<tr>
<td>BAN</td>
<td>0.43</td>
<td>0.47</td>
<td>0.39</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparisons of RMSE between BAN and others baselines for short-term and long-term motion sequence on 8 activities of the CMU dataset. The proposed BAN model consistently outperforms these baselines in almost all the scenarios.

<table>
<thead>
<tr>
<th>Motion</th>
<th>Tai chi</th>
<th>Dancing</th>
<th>Swordplay</th>
<th>Gymnastics</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing time</td>
<td>500</td>
<td>1000</td>
<td>1500</td>
<td>2000</td>
</tr>
<tr>
<td>SR-L1</td>
<td>1.93</td>
<td>1.98</td>
<td>2.23</td>
<td>2.47</td>
</tr>
<tr>
<td>LRMC</td>
<td>0.87</td>
<td>1.19</td>
<td>2.29</td>
<td>2.83</td>
</tr>
<tr>
<td>LSTM</td>
<td>1.35</td>
<td>1.45</td>
<td>1.28</td>
<td>1.34</td>
</tr>
<tr>
<td>FCNN</td>
<td>2.77</td>
<td>2.31</td>
<td>2.28</td>
<td>2.32</td>
</tr>
<tr>
<td>BAN</td>
<td>0.44</td>
<td>0.64</td>
<td>0.59</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2: Comparison of average bone-length error using different methods under different types of motion. The results of BAN are highlighted for each motion sequence.

5 Conclusion

In this work, we have proposed the bi-directional attention network, which can capture long-term dependency and motion correlation from forward and backward directions. This method effectively utilizes the spatio-temporal information of human motion by learning the relevant feature representation of each pose, which dramatically expands the performance of motion modeling. We demonstrate that our model significantly improves the performance of human motion recovery concerning accuracy and visualization results, even in the case of long sequences or different missing distributions. However, there are still two defects that cannot be ignored: High time consumption, because the LSTM encoding and attention weight computation are non-parallel; Performance degradation for handling high missing ratio (> 80%). Fortunately, the cases of high missing ratio rarely occur in real-world applications. In the future work, we plan to use the idea of the generative model to further expand the scope of application of the proposed model and consider applying it to other tasks of human motion.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO. 61772272) and the Project of Science and Technology of Jiangsu Province of China under Grant BE2017031.
References


