Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamic

Yongyi Tang[1*], **Lin Ma**[2†], **Wei Liu**[2], **Wei-Shi Zheng**[3†]
[1]School of Electronics and Information Technology, Sun Yat-sen University
[2]Tencent AI Lab
[3]School of Data and Computer Science, Sun Yat-sen University
{yongyi.tang92, forest.linma, wliu.cu}@gmail.com wszheng@ieee.org

Abstract

Human motion prediction aims at generating future frames of human motion based on an observed sequence of skeletons. Recent methods employ the latest hidden states of a recurrent neural network (RNN) to encode the historical skeletons, which can only address short-term prediction. In this work, we propose a motion context modeling by summarizing the historical human motion with respect to the current prediction. A modified highway unit (MHU) is proposed for efficiently eliminating motionless joints and estimating next pose given the motion context. Furthermore, we enhance the motion dynamic by minimizing the gram matrix loss for long-term motion prediction. Experimental results show that the proposed model can promisingly forecast the human future movements, which yields superior performances over related state-of-the-art approaches. Moreover, specifying the motion context with the activity labels enables our model to perform human motion transfer.

1 Introduction

Human motion prediction, serving as one of the most essential parts of robotic intelligence, enables rapid and high-fidelity reactions towards complex environment changes. For example, a robot can effortlessly avoid route collision by forecasting the movement about surrounding subjects. Nowadays, with the development of the MOCAP devices, such as Kinect, and the pose estimation algorithms [Yasin *et al.*, 2016; Tekin *et al.*, 2017], the sequence of human skeletons can be easily and accurately computed. It thus enables us to predict future human motion by analyzing the observed skeleton sequences, which can further help human action analysis/recognition, body pose estimation, and even human-robot interactions.

The historical human skeleton sequence needs to be effectively modeled for human motion prediction [Fragkiadaki *et al.*, 2015; Jain *et al.*, 2016] and action recognition [Wang *et al.*, 2014; Liu *et al.*, 2016]. Currently, one common strategy is to use a recurrent neural network (RNN) as the encoder along the temporal domain [Fragkiadaki *et al.*, 2015; Ghosh *et al.*, 2017; Martinez *et al.*, 2017], driven from sequence to sequence learning [Sutskever *et al.*, 2014], with the last hidden state encoding the motion context. For the existing recurrent units such as long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] and gated recurrent unit (GRU) [Cho *et al.*, 2014], the hidden states encode the skeleton sequence and update at every time step. Although LSTM and GRU are proposed to handle the long short-term depen-

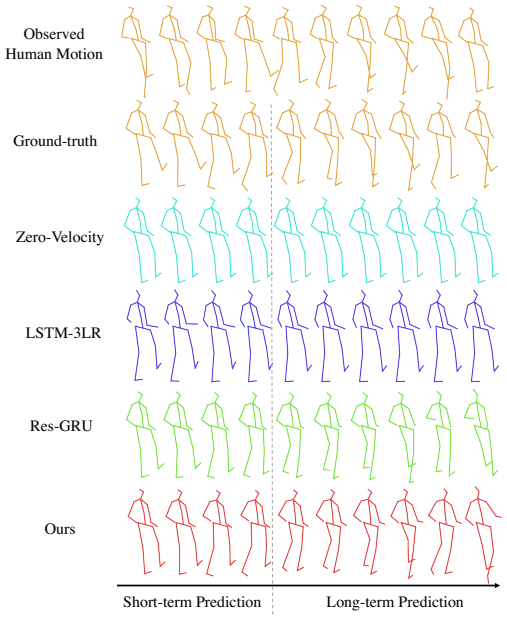

Figure 1: Human motion prediction on "walking". Top: the observed human motion sequence. Given the observed skeletons, the goal of this paper is to generate future skeletons similar to the ground-truth (the second row). Our method is able to well predict both short-term and long-term skeletons maintaining good temporal dynamic, while other existing methods fail to generate satisfying long-term skeletons. Better view in color.

[*]Work done while Yongyi Tang was a Research Intern with Tencent AI Lab.
[†]Corresponding authors.

dencies, the historical information, especially the long term one, cannot be well encoded with the updated hidden states overwhelmed by the input at current step [Bahdanau *et al.*, 2014]. Such information loss makes the long-term human motion prediction tend to converge to the mean pose or fail to produce motion dynamic, as the results of LSTM-3LR [Fragkiadaki *et al.*, 2015] and Res-GRU [Martinez *et al.*, 2017] shown in Fig. 1. Moreover, the human joints of skeleton are treated equally for the motion prediction in the prior works. Instead, the human motion can be viewed as the movement of the joints of the skeleton, where not every joint participates in the human pose evolutions. For the human activities, such as "walking" and "eating", the subject may stand still with the backbones motionless.

In order to make reliable future predictions, we model motion context by summarizing the historical human motion skeleton sequence with respect to the current skeleton. Such motion context can help to capture the human motion patterns, $i.e.$ the repeated patterns in "walking" and "eating", and ease the motion uncertainties, thus benefiting the long-term predictions. By utilizing both the pose information of the last frame and the summarized motion context, we propose a modified highway unit (MHU) to predict the future human skeleton. The MHU introduces a gate that can efficiently filter the motionless joints at each generation and pay more attentions on those with motion. Besides, in order to produce consistent human motions and enhance the motion dynamic, we introduce a gram matrix loss for minimization so as to explicitly penalize the mean pose convergence and ease error accumulation. These components enable our method to predict reliable long-term human motion as highlighted in the last row of Fig. 1.

In addition, prior works are only able to predict one single activity for a given pose sequence, $i.e.$ predicting future "walking" skeletons given the observed "walking" skeleton sequence. However, in realistic scenario, more than one type of activity may evolve given the observed human motion sequence. With our motion context modeling, we further exploit the ability of the proposed model on human motion transfer, which generates specific types of motion sequence given different action labels. As such, the human motion sequence can be manipulated by the given action labels, resulting in a smooth motion sequence with multiple activities, which will be detailed in Sec. 4.

Our contributions are summarized as follows: 1) We propose to model the motion context by summarizing the historical skeleton sequence with respect to the current one. MHU thereafter distinguishes the motion joints from the motionless ones to make effectively long-term human motion prediction. 2) A gram matrix loss is proposed for enhancing motion dynamic, which enables our model to produce highly correlated human motion in the temporal domain. 3) Our model can perform human motion transfer based on the motion context and the specified activity categories.

2 Related Works

Human Motion Analysis. Human motion analysis is one of the key problems in computer vision and robotics, and hence has received much attention [Aggarwal and Cai, 1997]. Human motion can be obtained by motion capture systems [Ionescu *et al.*, 2014] and Kinect device and extracted from videos [Brand and Hertzmann, 2000] and even static images [Li and Chan, 2014; Yasin *et al.*, 2016]. With the available body poses, several structural models such as hierarchical recurrent neural networks [Du *et al.*, 2015] and trust gates [Liu *et al.*, 2016] were proposed to address skeleton based action recognition. By representing skeletons with the rotation matrices, which forms a special orthogonal group SO(3), the researches in [Vemulapalli *et al.*, 2014] and [Huang *et al.*, 2017] developed group-based skeleton analysis for action recognition by using support vector machine and convolution neural network, respectively.

Human Motion Prediction. Human motion prediction aims to understand behaviors of a subject on the observed sequences and to generate future body poses. Deep learning based approaches have outperformed conventional methods on skeleton-based problem such as 3D pose estimation [Yasin *et al.*, 2016] and action recognition [Hu *et al.*, 2015; Liu *et al.*, 2016]. In this paper, we focus on human motion prediction based on deep neural networks. Prior works try to encode the observed information to latent variables and perform prediction as decoding by Restricted Boltzmann Machines (RBMs) [Taylor *et al.*, 2007]. [Fragkiadaki *et al.*, 2015] introduce Encoder-Recurrent-Decoder networks that learn the temporal dynamic of human motion by a long short-term memory (LSTM) model. They designed a non-linear transformation to encode pose feature and decode the output of the LSTM. The history information passes throughout the recurrent units to constrain human motion prediction. [Martinez *et al.*, 2017] further extended this scheme by modeling the velocity of joints instead of directly estimating the body pose, and employed single linear layer for pose features encoding and hidden states decoding. They find that the poses with zero-velocity achieve relatively less error on mean angle distance, which demonstrates the efficiency of the velocity modeling. To reduce the accumulated correlation error, a dropout auto-encoder (DAE) was proposed by [Ghosh *et al.*, 2017]. Apart from these approaches, structural RNN proposed by [Jain *et al.*, 2016] tries to capture the spatio-temporal relationship of joints.

3 Proposed Model

3.1 Problem Formulation

Given an observed sequence of body poses $\{\mathbf{x}^{t'}\}_{t'=1}^{T'}$ in 3D space, the goal of human motion prediction is to generate the consecutive human motion $\{\hat{\mathbf{x}}^{t}\}_{t=1}^{T}$ close to the ground-truths $\{\mathbf{x}^{t}\}_{t=1}^{T}$. Following the prior works [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017] for human motion prediction, the axis-angle representation of skeletons $\{\mathbf{x}^{t'}\}_{t'=1}^{T'}$ parameterizes a rotation of each joint in a three-dimensional Euclidean space by a rotation vector whose norm is the rotation angle.

Conventional RNN-based human motion prediction methods [Fragkiadaki *et al.*, 2015; Jain *et al.*, 2016; Martinez *et al.*, 2017] rely on the last hidden state $\mathbf{h}^{t-1}$ and predicted

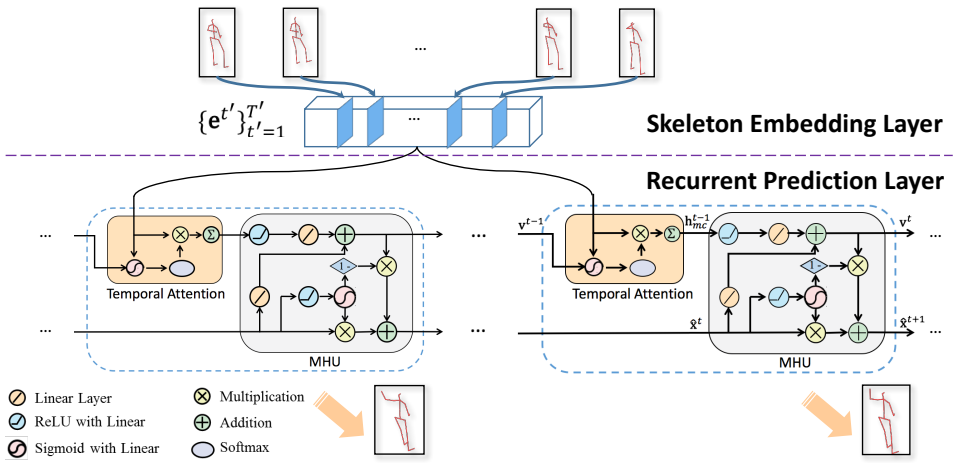

Figure 2: The architecture of our proposed model for human motion prediction. Each historical skeleton is first embedded into one semantic space. At each time step, the motion context modeling summarizes the skeleton embeddings with respect to the last predicted skeleton. Afterwards, MHU works on the motion context and the last estimated skeleton to yield the human motion at each time step.

skeleton $\hat{\mathbf{x}}^t$:

$$\hat{\mathbf{x}}^{t+1} = \text{RNN}(\hat{\mathbf{x}}^t, \mathbf{h}^{t-1}). \quad (1)$$

The historical human skeletons are encoded by $\mathbf{h}^0$ for predicting the first skeleton $\hat{\mathbf{x}}^1$. However, the failure cases in long-term human motion prediction (as shown in Fig. 1) indicate that using the final hidden state as the motion context is not satisfactory to well capture the historical motion information.

To address the above problems, we aim at designing a model f equipped with motion context modeling to fully explore the properties of human motion sequences, which is further formulated as:

$$\hat{\mathbf{x}}^{t+1} = f(\hat{\mathbf{x}}^t, \{\mathbf{x}^{t'}\}_{t'=1}^{T'}). \quad (2)$$

Our proposed model f directly accesses to the historical human skeletons $\{\mathbf{x}^{t'}\}_{t'=1}^{T'}$ at each step for prediction, which enables us to yield a more representative motion context. As such, the model can simply repeat the observed pattern to get a reasonable prediction for periodic activities such as "walking" and "eating". For aperiodic activities, the encoded motion context can still provide meaningful information of historical activities (such as directions or the habit of movement), and thus further reduce the search space for making predictions.

3.2 Our Approach

As shown in Fig.2, the proposed model mainly consists of two components: a skeleton embedding layer and a recurrent prediction layer. The embedding layer can be regarded as an encoder, and the recurrent prediction layer is thus denoted as the decoder, which consists of two main components, namely the motion context modeling and the modified highway unit (MHU). These two components are coupled together, and this enables the proposed framework to predict reliable long-term human motions.

A multi-layer non-linear network is constructed to realize the skeleton embedding layer, which projects the observed skeletons $\{\mathbf{x}^{t'}\}_{t'=1}^{T'}$ into the semantic space yielding $\{\mathbf{e}^{t'}\}_{t'=1}^{T'}$. Specifically, we concatenate the output of a fully connected layer $\mathbf{h}_{e1} = \mathbf{W}_{e1}\mathbf{x}^{t'} + \mathbf{b}_{e1}$ and its activated output $\mathbf{h}_{e2} = \text{ReLU}(\mathbf{h}_{e1})$, and finally preform embedding by $\mathbf{e}^{t'} = \mathbf{W}_{e2}[\mathbf{h}_{e1}; \mathbf{h}_{e2}] + \mathbf{b}_{e2}$.

During the prediction, the motion context is firstly summarized from the skeleton embeddings with respect to the last predicted skeleton. Afterwards, the MHU exploits the relationships between the motion context and the predicted skeleton to generate the human motion at each time step.

Motion Context Modeling

Motion context modeling aims at encoding the historical human motion, which can further boost the future skeleton prediction. Existing methods model motion context simply by LSTM or GRU, and the last hidden state is taken as motion context for human motion prediction [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017; Zimo *et al.*, 2017] and action recognition [Liu *et al.*, 2016; Du *et al.*, 2015]. However, the last hidden state in RNN is usually dominated by the input at the latest time step. Therefore, the previous information, especially for the long-term one, is not effectively encoded into the hidden state. While for future motion prediction, the historical skeletons are believed to be helpful.

In this paper, we propose to use temporal attention mechanism [Bahdanau *et al.*, 2014] to summarize all the historical skeletons with the respect to predicted one at each time step:

$$\beta^{t'} = \mathbf{W}_\beta \tanh(\mathbf{U}_{\beta v}\mathbf{v}^{t-1} + \mathbf{U}_{\beta e}\mathbf{e}^{t'} + b_\beta), \quad (3)$$

$$\alpha^{t'} = \frac{\exp(\beta^{t'})}{\sum_{t'=1}^{T'} \exp(\beta^{t'})}, \quad (4)$$

where $\mathbf{v}^{t-1}$ denotes the predicted skeleton at time $t-1$. $\alpha^{t'}$ denotes the attentive weight with respect to each historical skeleton. The temporal attention mechanism directly works on the skeleton embeddings, which can more effectively capture the relations between the predicted skeleton and historical motion. With the computed attentive weights, the motion

context is thus computed by:

$$\mathbf{h}_{mc}^{t-1} = \sum_{t'=1}^{T'} \alpha^{t'} \mathbf{e}^{t'}. \quad (5)$$

The obtained motion context $\mathbf{h}_{mc}^{t-1}$ can selectively summarize the historical skeleton information. The obtained motion context presents no bias on short-term or long-term information. Thus, $\mathbf{h}_{mc}^{t-1}$ can help produce more reliable long-term predictions compared with the state-of-the-art methods which directly use the last hidden state of traditional RNNs.

Modified Highway Unit

The human motion can be viewed as the movement of skeleton joints, where not every joint participates in pose evolutions. For example, one subject mainly stands still with the backbones presenting motionless in activities such as "phoning" and "eating". Therefore, the human motion is only triggered by the activity-specific skeleton joints.

Based on these observations, we introduce an MHU in our recurrent prediction layer as shown in Fig.2 in order to efficiently model the skeleton joints that contain meaningful motion information. We introduce ReLU non-linearity in the Recurrent Highway Network proposed by [Zilly *et al.*, 2016] for both skeleton estimation and gate estimation. We additionally drop the $tanh$ activation in the vanilla RHN.

Given the current input skeleton representation $\mathbf{x}^t$ and the motion context $\mathbf{h}_{mc}^{t-1}$ from the last time step, our proposed MHU is formulated as:

$$\mathbf{v}^t = \mathbf{W}_v \phi(\mathbf{U}_{vh}\mathbf{h}_{mc}^{t-1} + \mathbf{b}_v) + \mathbf{U}_{vx}\mathbf{x}^t + \mathbf{b}_{vh}, \quad (6)$$

$$\mathbf{z}^t = \sigma(\mathbf{W}_z \phi(\mathbf{U}_{zx}\mathbf{x}^t + \mathbf{b}_z) + \mathbf{b}_{zx}), \quad (7)$$

$$\hat{\mathbf{x}}^{t+1} = (\mathbf{1} - \mathbf{z}^t) \odot \mathbf{v}^t + \mathbf{z}^t \odot \mathbf{x}^t, \quad (8)$$

where $\odot$ is element-wise multiplication, $\mathbf{W}$, $\mathbf{U}$ and b are the learned parameters, and ϕ and σ denote the rectified linear unit and sigmoid function, respectively.

Note that $\mathbf{z}^t$ ranges in [0,1] for gating the current skeleton $\mathbf{x}^t$ and the estimated next joints $\mathbf{v}^t$. The modeling of gate state $\mathbf{z}^t$ involves a non-linear transformation, which implicitly captures the spatial relations of $\mathbf{x}^t$. As such, MHU is expected to focus on the joints with large motions, and conducts partially updating in Eq.8 by the element-wise multiplication. These non-linear operations within the MHU can help explore spatial relations of skeleton joints.

3.3 Enhancing Motion Dynamics with Gram Matrix Objective

In addition to model motion context for long-term prediction, the transitions between skeletons should be addressed to generate dynamic human motion and prevent mean pose convergence. To minimize the error of human motion prediction, existing methods [Fragkiadaki *et al.*, 2015; Jain *et al.*, 2016; Martinez *et al.*, 2017] usually adopt mean square error (MSE) as the objective function. However, the MSE constrains models to generate the human motion that stays around the center of the ground-truth distribution, which are the mean poses. Moreover, the MSE objective only treats each motion independently which may cause motion inconsistency. Instead, we propose to minimize the gram matrix between consecutive motions, which is defined as follow:

$$\mathcal{L}_{gram} = \frac{1}{T}\sum_{t=1}^{T-1} \left|\left|G(\hat{\mathbf{x}}^t, \hat{\mathbf{x}}^{t-1}) - G(\mathbf{x}^t, \mathbf{x}^{t-1})\right|\right|_2^2, \quad (9)$$

where the gram matrix $G(\mathbf{x}^t, \mathbf{x}^{t-1})$ is defined as:

$$G(\mathbf{x}^t, \mathbf{x}^{t-1}) = [\mathbf{x}^t; \mathbf{x}^{t-1}][\mathbf{x}^t; \mathbf{x}^{t-1}]^\top, \quad (10)$$

and $[\cdot;\cdot]$ denotes the concatenation of vectors.

On one hand, the correlation between skeleton joints is represented in the gram matrix such that the spatial relation among different skeleton joints can be further explored. On the other hand, the temporal dynamic is captured by the correlation between $\mathbf{x}^t$ and $\mathbf{x}^{t+1}$, which enables our model to enhance human motion along temporal axis. For the action such as "walking", the arms and legs move alternately. Such spatial-temporal correlation represented in the gram matrix can enable producing human-like walking motion. Thus both short-term and long-term human motion predictions can be improved.

4 Experiments

4.1 Experimental Settings

H3.6m Mocap Dataset for Human Motion Prediction. We conducted our experiments of human motion prediction on the H3.6m mocap Dataset [Ionescu *et al.*, 2014], which is the largest human motion dataset for 3D body pose analysis. It consists of 15 activities including periodic activities like "walking" and non-periodic activities such as "discussion" and "taking photo", performed by seven different professional actors. Recorded by a Vicon motion capture system, the H3.6m dataset provides high quality 3D body joint locations in the global coordinate sampled at 50 frames per second (fps).

Data Representation and Preprocessing. For all our experiments, we followed the same data setting in [Fragkiadaki *et al.*, 2015; Jain *et al.*, 2016; Martinez *et al.*, 2017]. The motion sequence was down-sampled by 2 to 25 fps. And 5 subjects were selected for testing with the others for training. The joint features were represented in exponential map [Grassia, 1998] which is also known as the angle-axis representation. The three dimension feature of each joint represents the rotation vector with respect to the parent joint predefined in H3.6m dataset. All the features were normalized into the range of [-1,1]. We did not use the label as additional information except for the experiments of human motion transfer.

Training. Single layer of MHU with 1024 units was adopted in all our experiments. Empirically, stacking more layers of MHU did not help improve the performance. To better capture human motion, all the activities were trained together for prediction as a default setting. We used $T' = 30$ observed frames for embedding to estimate future $T = 10$ frames. We used stochastic gradient descent with the momentum setting to 0.9. The learning rate was set to 0.05 decayed

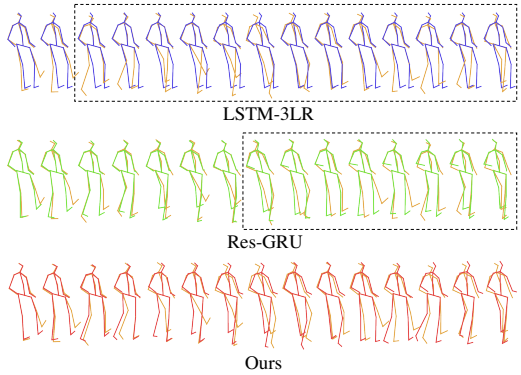

Figure 3: The comparison of mean pose convergence of the walking activity. The ground-truth poses are shown in yellow. The dash boxes highlight the converging motion sequence.

with factor of 0.95 for every 10,000 steps. And the gradient was clipped to a maximum L2-norm of 5. Batch size of 80 was used throughout our experiments. Normally, the training converged in around 20,000 steps.

4.2 Experimental Results

We first evaluate the ability of the proposed framework for predicting human motion and made comparison with related recent state-of-the-art methods including ERD [Fragkiadaki *et al.*, 2015], LSTM-3LR [Jain *et al.*, 2016] and Res-GRU [Martinez *et al.*, 2017]. We reproduced the results of these methods. Please note that our reproduced results often present better performance than that reported in their papers. Following the evaluation of previous works, we converted the representation of joints from angle-axis to angle of rotation, and thereby measured the Euclidean distance between the predicted joints and its ground-truth by:

$$D(\{\mathbf{x}^t\}_{t=1}^T, \{\hat{\mathbf{x}}^t\}_{t=1}^T) = \sum_{t=1}^{T}\sum_{i=1}^{N}\sqrt{d^2(\alpha_i^t, \hat{\alpha}_i^t) + d^2(\beta_i^t, \hat{\beta}_i^t) + d^2(\gamma_i^t, \hat{\gamma}_i^t)}, \quad (11)$$

where $d(a, b) = \min\{|a - b|, 2\pi - |a - b|\}$.

Interestingly, the research in [Martinez *et al.*, 2017] found that repeating the last body pose also gave a relative small error in the measurement of the Euclidean distance between the ground-truth, which performed even better than ERD [Fragkiadaki *et al.*, 2015] and LSTM-3LR [Fragkiadaki *et al.*, 2015]. One possible reason is that the human motion within the dataset is slight for some activities. Therefore, simply repeating the last body pose can yield the reasonable objective results. Another possible reason may be attributed to the evaluation metric, which is an Euclidean distance and can only depict independent distance for each joint, and thus this ignores the relations between joints. Thus even with a smaller Euclidean distance, the motion prediction may not be plausible.

We compare our results with the existing methods and the variants of our method. The first variant is using conventional mean square loss as in [Martinez *et al.*, 2017; Fragkiadaki *et al.*, 2015] and encoder-decoder framework with MHU. We then replaced the MSE training loss with the gram matrix loss in the second variant. The fist and second variations are named as "MHU-MSE" and "MHU-Gram", respectively.

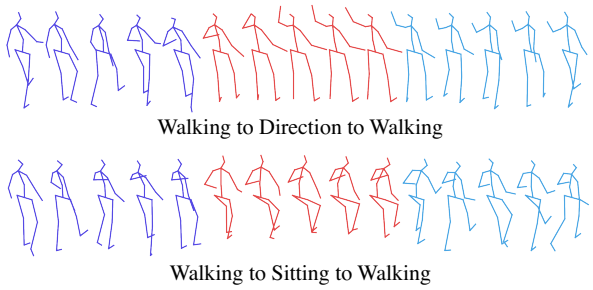

Figure 4: The result of human motion transfer between two different activities. The observed pose sequences are shown in purple, and the predicted motion are shown in red and blue. Better view in color.

The overall human motion prediction result of all 15 activities of H3.6m dataset via mean angle error is shown in Table.1. It can be observed that our methods as well as two different variants can outperform the competitors. More specifically, the result shows that MHU-MSE performs well in the short-term improvement of prediction especially from 160ms to 720ms. The reason can be attributed to that the MHU can efficiently filter motionless joints and propagate information between two layers with the modified non-linearity. For the very short-term prediction, the spatial information of the body pose predominates the measurement. For longer term prediction, the motion information is more important. Therefore, only considering the spatial pose information cannot well model the motion dynamic. As such MSE-Gram, which targets at enhancing motion dynamic, performs better, which achieves 1.82 on mean angle error at 1000ms. Finally, by assembling the motion context modeling, our model can achieve the best performance on both short-term and long-term predictions.

In details, we show a part of the results in Table. 2 which contains both short-term and long-term comparisons with the compared methods. In most of the cases, our results are competitive in short-term prediction, and clearly outperform the baseline methods in long-term prediction. For the "walking" activity, the objective evaluation of our result is close to that of Res-GRU. However, by visualizing the motion in Fig. 3, it seems that the results generated by LSTM-3LR method and Res-GRU method converge to the mean body pose. On the contrary, our method can resemble the "walking" behaviors of the body pose, thus presenting the predicted motion with highly dynamic.

4.3 Human Motion Transfer

As mentioned before, the long-term human motion is not deterministic and may alter according to subjective factors. For example, while one is sitting, at any time he/she may suddenly stand up and walk around. Here, we try to simulate this situation by modifying the hidden state of each decoder input

Methods	Short-term					Long-term							
	80ms	160ms	240ms	320ms	400ms	480ms	560ms	640ms	720ms	800ms	880ms	960ms	1000ms
ERD [Fragkiadaki *et al.*, 2015]	0.93	1.07	1.19	1.31	1.41	1.52	1.58	1.64	1.70	1.78	1.86	1.93	1.95
LSTM-3LR [Fragkiadaki *et al.*, 2015]	0.87	0.93	1.06	1.19	1.30	1.41	1.49	1.55	1.62	1.70	1.79	1.86	1.89
Res-GRU [Martinez *et al.*, 2017]	0.40	0.72	0.92	1.09	1.23	1.36	1.45	1.52	1.59	1.68	1.77	1.85	1.89
Zero-velocity	0.40	0.71	0.90	1.07	1.20	1.32	1.42	1.50	1.57	1.66	1.75	1.82	1.85
MHU-MSE	**0.39**	0.69	0.88	1.04	1.17	1.30	1.40	1.49	1.57	1.67	1.77	1.86	1.89
MHU-Gram	**0.39**	**0.68**	0.86	**1.01**	1.14	1.26	1.35	1.43	1.50	1.60	1.70	1.79	1.82
Ours	**0.39**	**0.68**	**0.85**	**1.01**	**1.13**	**1.25**	**1.34**	**1.42**	**1.49**	**1.59**	**1.69**	**1.77**	**1.80**

Table 1: Performance comparison between different methods in terms of both short-term and long-term human motion prediction of all 15 activities from H3.6m dataset via mean angle error. The best performance is highlighted in boldface.

Methods	Walking								Greeting							
	Short-term				Long-term				Short-term				Long-term			
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD [Fragkiadaki *et al.*, 2015]	0.77	0.90	1.12	1.25	1.44	1.45	1.46	1.44	0.85	1.09	1.45	1.64	1.93	1.89	1.92	1.98
LSTM-3LR [Fragkiadaki *et al.*, 2015]	0.73	0.81	1.05	1.18	1.34	1.36	1.37	1.36	0.80	0.99	1.37	1.54	1.81	1.76	1.79	1.85
Res-GRU [Martinez *et al.*, 2017]	**0.27**	**0.47**	**0.68**	**0.76**	**0.90**	**0.94**	0.99	**1.06**	**0.52**	**0.86**	1.30	1.47	1.78	1.75	1.82	1.96
Zero-velocity	0.39	0.68	0.99	1.15	1.35	1.37	1.37	1.32	0.54	0.89	1.30	1.49	1.79	1.74	1.77	**1.80**
Ours	0.32	0.53	0.69	0.77	**0.90**	**0.94**	**0.97**	**1.06**	0.54	0.87	**1.27**	**1.45**	**1.75**	**1.71**	**1.74**	1.87
Methods	Walking Dog								Discussion							
	Short-term				Long-term				Short-term				Long-term			
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD [Fragkiadaki *et al.*, 2015]	0.99	1.25	1.48	1.58	1.83	1.88	1.96	2.03	0.76	0.96	1.17	1.24	1.57	1.70	1.84	2.04
LSTM-3LR [Fragkiadaki *et al.*, 2015]	0.91	1.07	1.39	1.53	1.81	1.85	1.90	2.00	0.71	0.84	1.02	1.11	1.49	1.62	1.76	1.99
Res-GRU [Martinez *et al.*, 2017]	**0.56**	0.95	1.33	1.48	1.78	1.81	1.88	1.96	**0.31**	0.69	1.03	1.12	1.52	1.61	1.70	**1.87**
Zero-velocity	0.60	0.98	1.36	1.50	1.74	1.80	1.87	1.96	**0.31**	0.67	0.97	1.04	1.41	1.56	1.71	1.96
Ours	**0.56**	**0.88**	**1.21**	**1.37**	**1.67**	**1.72**	**1.81**	**1.90**	**0.31**	**0.66**	**0.93**	**1.00**	**1.37**	**1.51**	**1.66**	1.88
Methods	Posting								Taking Photo							
	Short-term				Long-term				Short-term				Long-term			
	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1000ms
ERD[Fragkiadaki *et al.*, 2015]	1.13	1.20	1.59	1.78	1.86	2.03	2.09	2.59	0.70	0.78	0.97	1.09	1.20	1.23	1.27	1.37
LSTM-3LR[Fragkiadaki *et al.*, 2015]	1.08	1.01	1.42	1.61	1.79	2.07	2.13	2.66	0.63	0.64	0.86	0.98	1.09	1.13	1.17	1.30
Res-GRU[Martinez *et al.*, 2017]	0.41	0.84	1.53	1.81	2.06	2.21	2.24	2.53	0.29	0.58	0.90	1.04	1.17	1.23	1.29	1.47
Zero-velocity	**0.28**	**0.57**	**1.13**	**1.37**	**1.81**	2.14	2.23	2.78	**0.25**	**0.51**	**0.79**	**0.92**	**1.03**	**1.06**	**1.13**	**1.27**
Ours	0.33	0.64	1.22	1.47	1.82	**2.11**	**2.17**	**2.51**	0.27	0.54	0.84	0.96	1.04	1.08	1.14	1.35

Table 2: Performance comparison between different methods in terms of both short-term and long-term human motion prediction via mean angle error for each individual activity from H3.6m dataset, including "`walking`", "`Greeting`", "`Walking Dog`", "`Discussion`", "`Posting`" and "`Taking Photo`". The best performance is highlighted in boldface.

with the embedded activity label different from the observed activity.

Since on the H3.6m dataset, there is no ground-truth human motion sequence that contains two different activities, we thus trained our model with the activity labels and feed a specific label at the test time. More specifically, after processing temporal attention procedure, we concatenated the hidden state with embedded action label, which was fed into a nonlinear layer in order to construct the context representation for the decoding.

We show four examples of human motion transfer in Fig. 4 including "walking" to "direction", "sitting" to "walking" and the inverses. By modifying the hidden states that encode the motion context of the subject, our method is able to transfer[1] human motion with smooth activity transitions. In details, as shown in the third row of Fig. 4, the proposed model manages to transfer human motion from "walking" to "sitting" and the inverse motion. This result shows that the motion context encoded in the hidden states can be well extracted by the designed MHU to produce reliable human motion transfer.

[1]We use 'transfer' here because there is no ground-truth sequence to be 'predicted'.

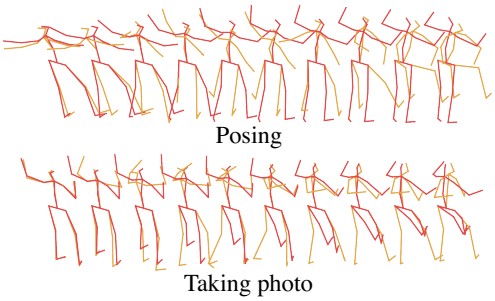

Figure 5: Failure cases of our human motion prediction method. The ground-truths and our results are shown in yellow and red, respectively.

4.4 Limitations

Fig. 5 illustrates some failure cases of our method, which also happen for the existing methods. The main reason is that these activities are of high uncertainty with different subjects. Therefore, the observed information cannot provide enough evidence for modeling and predicting.

5 Conclusion

In this paper, we have proposed a new model to predict long-term human motions by exploring motion context and en-

hancing motion dynamic. The proposed motion context summarized the historical skeletons for providing fully observed evidence in long-term prediction. To enhance motion dynamic, the gram matrix training loss is further incorporated to capture the temporal transitions. The extensive results demonstrate that our proposed model outperforms existing methods especially for long-term motion prediction. Moreover, compared with other models, our model can perform human motion transfer which makes motion prediction based on the action command and alters the generated motion types accordingly.

Acknowledgments

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(61522115, 61661130157), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), and the Royal Society Newton Advanced Fellowship (NA150459).

References

[Aggarwal and Cai, 1997] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102. IEEE, 1997.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Brand and Hertzmann, 2000] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192. ACM Press/Addison-Wesley Publishing Co., 2000.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Du *et al.*, 2015] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[Fragkiadaki *et al.*, 2015] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[Ghosh *et al.*, 2017] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. *arXiv preprint arXiv:1704.02827*, 2017.

[Grassia, 1998] F Sebastian Grassia. Practical parameterization of rotations using the exponential map. *Journal of graphics tools*, 3(3):29–48, 1998.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hu *et al.*, 2015] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[Huang *et al.*, 2017] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6099–6108. IEEE computer Society, 2017.

[Ionescu *et al.*, 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[Jain *et al.*, 2016] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[Li and Chan, 2014] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[Liu *et al.*, 2016] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[Martinez *et al.*, 2017] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. *arXiv preprint arXiv:1705.02445*, 2017.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[Taylor *et al.*, 2007] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2007.

[Tekin *et al.*, 2017] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *International Conference on Computer Vision (ICCV)*, number EPFL-CONF-230311, 2017.

[Vemulapalli *et al.*, 2014] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

[Wang *et al.*, 2014] Jiang Wang, Zicheng Liu, and Ying Wu. Learning actionlet ensemble for 3d human action recognition. In *Human Action Recognition with Depth Cameras*, pages 11–40. Springer, 2014.

[Yasin *et al.*, 2016] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.

[Zilly *et al.*, 2016] Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. Recurrent highway networks. *arXiv preprint arXiv:1607.03474*, 2016.

[Zimo *et al.*, 2017] Li Zimo, Yi Xiao, He Chong, and Li Hao. Auto-conditioned lstm network for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017.