On Retrospecting Human Dynamics with Attention

Minjing Dong and **Chang Xu**
School of Computer Science, Faculty of Engineering, University of Sydney, Australia
mdon0736@uni.sydney.edu.au, c.xu@sydney.edu.au

Abstract

Deep recurrent neural networks have achieved impressive success in forecasting human motion with a sequence to sequence architecture. However, forecasting in longer time horizons often leads to implausible human poses or converges to mean poses, because of error accumulation and difficulties in keeping track of longerterm information. To address these challenges, we propose to retrospect human dynamics with attention. A retrospection module is designed upon RNN to regularly retrospect past frames and correct mistakes in time. This significantly improves the memory of RNN and provides sufficient information for the decoder networks to generate longer term prediction. Moreover, we present a spatial attention module to explore and exploit cooperation among joints in performing a particular motion. Residual connections are also included to guarantee the performance of short term prediction. We evaluate the proposed algorithm on the largest and most challenging Human 3.6M dataset in the field. Experimental results demonstrate the necessity of investigating motion prediction in a self audit manner and the effectiveness of the proposed algorithm in both short term and long term predictions.

1 Introduction

Human dynamics modelling has received increasing attention in recent years, considering its wide application in different scenarios, such as autonomous driving systems and human-robot interactions. The target of human motion prediction is to generate future continuous and realistic human poses given a seed sequence, which can further assist human motion analysis. For example, forecasting the motion of pedestrian is essential for self-driving cars to avoid collision, and anticipating human motion could boost the understanding of user intent for a seamless human-machine collaborations.

Human motions in practice can be rather complicated, and often of high uncertainty, which makes the human motion prediction task difficult and challenging. Thanks to the development of human motion capture systems and pose estimation algorithms [Nie *et al.*, 2018; Tekin *et al.*, 2016], large-scale human motion datasets are available for investigating machine learning approaches in human dynamics analysis. Traditional approaches focus on hidden Markov models [Lehrmann *et al.*, 2014], linear dynamical systems [Pavlovic *et al.*, 2000], Gaussian process latent variable models [Urtasun *et al.*, 2008; Xu *et al.*, 2013] and bilinear spatio-temporal basis models [Akhter *et al.*, 2012]. However, there exists trade-offs between model capacity and inference complexity for these approaches, which make them difficult to be trained on large datasets. Motivated by the success of Recurrent Neural Networks (RNN), a wide variety of RNN-based methods have emerged to tackle the human motion analysis problem. For example, Encoder-Recurrent-Decoder network [Fragkiadaki *et al.*, 2015] has been proposed to learn temporal dependencies, with spatial encoder and decoder wrapped upon recurrent cell and last hidden state encoding human poses. Besides directly encoding human poses by hidden variables, Residual RNN models velocity representations, which boosts the short-term prediction performance by applying seq2seq model with residual connections [Martinez *et al.*, 2017].

Though these methods have achieved impressive performance in analyzing human motion, there are still some drawbacks. Firstly from the temporal aspect, these RNN-based methods are devoted to predict future sequence and keep moving forward without looking back to handle error accumulation, which increases the difficulties in maintaining faraway information. On the other hand, considering physical limitations (e.g. gravity) and structural constraints between body parts, it is essential to highlight different importance of body joints in motion. Given an action across different frames, each joint would have distinct levels of movements and those with more movements deserve more attention. Given these drawbacks, it is difficult for existing approaches to accomplish plausible motion predictions for aperiodic actions, especially over long time horizon [Gehring *et al.*, 2017], due to the mean pose problem. To deal with such a limitation, GAN networks are also proposed, [Barsoum *et al.*, 2017] is the first to do probabilistic motion prediction with WGAN-GP. Although different losses have been added, it's still hard to tell if the training has converged. Fidelity and continuity discriminators with geodesic loss introduced by [Gui *et al.*, 2018] to boost the performance, however, it uses action labels as input, which makes it supervised method.

In this paper, we propose to retrospect human dynamics with attention. It is necessary to look back in order to move forward. Accounting for past mistakes not only enables a timely self-correction, but also strengthens the confidence in making the subsequent prediction. We develop a retrospection module to constantly examine past motion sequence, which is beneficial for keeping track of longer-term motion information. A spatial attention module is designed to distinguish importance of different joints in each frame, so that more attention can be paid on joints that involve more movements or movement tendency. The proposed algorithm outperforms RRNN in both quantitative and qualitative evaluations on the Human 3.6M dataset. We can produces more realistic and coherent human motion predictions.

2 Methodology

We adapt sequence-to-sequence (seq2seq) architecture [Sutskever *et al.*, 2014], which is widely used in recent RNN based methods for motion generation, as shown in Figure 1(a). It consists of two networks, an encoder which takes as input a sequence of observed human poses and a decoder. The encoder takes as input a sequence of observed human poses and generates latent representations. The decoder produces the predicted poses according to the latent representations.

Formally consider an observed seed sequence of human poses $X_{1:t} = [x_1, x_2, ..., x_t]$, where $x_i \in \mathbb{R}^K$ is the representation of skeletons corresponding to a particular human pose and K is the number of joint angles. The objective of human motion prediction is to produce the continuous human poses after $X_{1:t}$, noted as $\hat{X}_{(t+1):(t+T)}$, which are close to ground truth $X_{(t+1):(t+T)}$, where T is the length of prediction sequence. The historical information can be maintained by Gated recurrent unit (GRU) [Cho *et al.*, 2014] by keep updating its hidden state at each time step. Thus, we have a sequence of hidden states $h_{1:t+T-1}$. The traditional objective in this task is to minimize the mean squared error (MSE) between the ground truth and prediction sequence as

$$\mathcal{L}_{seq} = \frac{1}{T} \sum_{t'=t+1}^{t+T} ||\hat{x}_t - x_t||_2^2. \tag{1}$$

It is difficult for RNN based methods to keep track of long-term information and capture spatial correlations, which would cause larger error and generate static or even unrealistic poses [Li *et al.*, 2018]. To handle long-term dependencies and capture spatial dynamics accurately, we propose a retrospection module with attention upon GRU. Residual connection [Martinez *et al.*, 2017] is deployed to enable the decoder to learn velocity representation instead of human poses directly, which improves short-term predictions and motion continuity. For the spatial decoder network wrapper upon GRU cell, we use two fully-connected layers with dropout to prevent overfitting and further explore the spatial correlations. An overview of the proposed algorithm for human motion prediction is shown in Figure 1.

2.1 Retrospection Module

We construct a retrospection module (RM), which can be regarded as a temporary memory to retrospect previous information. Equipped with attention techniques, RM can assist GRU to memorize long-term information as well as capture temporal correlations. Since human motions are continuous, complicated, and always of high uncertainty, the performance of motion predictions can highly depend on many frames in the sequence. As a result, the retrospection module shall be executed for several times over the whole sequence to retrospect sufficient information. To accomplish this, we set anchor points every C frames on GRU's hidden states as

$$P_{1:n} = h_{1::C} = [h_C, h_{2C}, ..., h_{nC}], \tag{2}$$

where n is the number of anchor points, and nC is less than $t+T$. Figure 1(c) illustrates the architecture of our retrospection module.

We set a retrospection module at each anchor point. In particular, we select a subsequence before the anchor point, and feed RM the first token of this subsequence and the hidden state of this anchor point to initialize the retrospection. The decoder network in RM is expected to predict the rest of this subsequence. The decoder in RM shares the same weights with that in seq2seq, as shown in Figure 1(a). For expression simplicity, we adopt a new variable y to represent elements in the sequence $\{x_1, \cdots, x_t, \hat{x}_{t+1}, \cdots, \hat{x}_{t+T}\}$. Given the k-th anchor point, we select the subsequence $Y = \{y_{(k-1)C+1}, \cdots, y_{kC}\}$, where the length of the subsequence has been fixed as C. The hidden state is calculated as,

$$\hat{h}_{s+1} = GRU(y_s, P_k), \tag{3}$$

where $s = (k-1)C+1$, function GRU denotes one step update of GRU cell and P_k is the hidden state corresponding to the k-th anchor point, which is the initial hidden state. Then the first human pose generated through RM is computed as

$$\hat{y}_{s+1} = f(\hat{h}_{s+1}) + y_s, \tag{4}$$

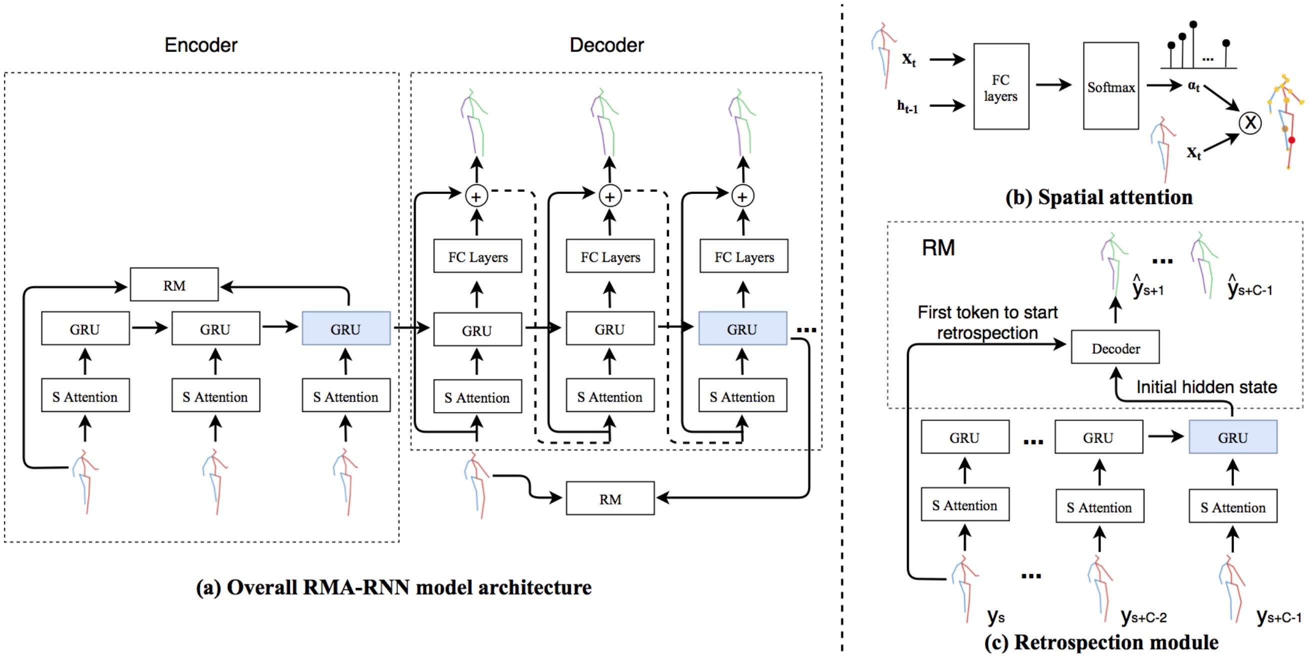
where function f represents two fully-connected layers forward operation. According to [Martinez *et al.*, 2017], residual connection is also adopted in Eq. (4), so that $f(\hat{h}_{s+1})$ can represent the velocity and $f(\hat{h}_{s+1}) + y_s$ represents the output human pose. Taking $\hat{y}_{s+1}$ as the new y_s and $\hat{h}_{s+1}$ as the new P_k in Eq. (3), the hidden state in GRU cell can be updated, and we can easily predict the next frame $\hat{y}_{s+2}$ in Eq. (4) accordingly. The rest subsequence can thus be predicted recursively by repeating the aforementioned calculations.

Note that y_s is the first token of subsequence to initialize the prediction, which is not part of the predicted subsequence, the length of predicted one is $C-1$. Given this predicted subsequence $\{\hat{y}_{s+1}, \cdots, \hat{y}_{s+C-1}\}$, the predicted loss for the k-th retrospect module can be computed as

$$\mathcal{L}_{RM}(k) = \frac{1}{C-1} \sum_{t'=s+1}^{s+C-1} ||\hat{y}_{t'} - x_{t'}||_2^2. \tag{5}$$

Note that for each anchor point, we need to predict $C-1$ frames before it. The overall loss for the whole sequence $\{x_1, \cdots, x_t, \hat{x}_{t+1}, \cdots, \hat{x}_{t+T}\}$ can then be written as

$$\mathcal{L}_{RM} = \sum_{k=1}^{n} \mathcal{L}_{RM}(k). \tag{6}$$



(a) Overall RMA-RNN model architecture

(b) Spatial attention

(c) Retrospection module

Figure 1: An illustration of our retrospection module with attention on RNN model (RMA-RNN). The red-blue skeletons represent the ground truth, and the green-purple skeletons represent the prediction. GRU cells with blue background are anchor points selected, which are initial states of retrospection module (RM). We introduce spatial attention to focus on joints with more movements (Fig. (b)). With the current GRU hidden state and the first token of subsequence before anchor points, retrospection module predicts the rest of subsequence for each anchor point, as a process of looking back previous frames (Fig. (c)).

Finally, we combine this retrospection loss with the original seq2seq loss to achieve the resulting objective function:

$$\mathcal{L}_{model} = \mathcal{L}_{seq} + \alpha \mathcal{L}_{RM}, \quad (7)$$

where α is a hyper parameter to balance the influence of these two terms. The proposed retrospection module can significantly enhance the ordinary RRNN method by fully investigating previous information. By retrospecting subsequence before the anchor point, short-term dependencies in frames can be captured. On the other hand, multiple anchor points have been set over the entire sequence to prevent the encoded information from vanishing and improve long-term memory.

2.2 Spatial Attention

Attention mechanism is helpful in various tasks [Chen *et al.*, 2018]. In this paper, we propose to explore different importance of joints by assigning spatial attention weights on all the joint angles, so that more attention could be paid on the joint angles that are more informative in describing the motion. For example, "walking" involves more movement on leg joints, however, "smoking" involves more rotation of arms.

Figure 1(b) illustrates the framework of the proposed spatial attention module. We suppose that the importance of joints mainly depends on the current input pose and the last hidden state that represents the motion velocity. At each time step t, given the input pose $x_t = [x_{t,1}, x_{t,2}, ..., x_{t,K}]$, the attention scores are computed as

$$score(x_t) = W_a \tanh(W_x x_t + W_h h_{t-1} + b_{xh}) + b_a, \quad (8)$$

where h_{t-1} is the last hidden state, W_a, W_x and W_h are weight matrices and b_{xh} and b_a are bias vectors. This score stands for the importance of each joint angle and is normalized by a Softmax layer as

$$a_{t,n} = \frac{exp(score(x_{t,n}))}{\sum_{i=1}^{K} exp(score(x_{t,i}))}, \quad (9)$$

where $n \in [1, K]$. Instead of taking original input x_t, the modified input $a_t \cdot x_t$ using spatial attention would be more beneficial for the subsequent processing, as the informative joint angles can be highlighted while those minor ones are weaken in the computation. GRU cells are adopted after the spatial attention module to further process the pose vectors. Thus, the inputs fed to GRU cell in Eq. (3), y_s and $\hat{y}_{s+i}$ ($i \in [1, C-1]$), will be replaced with $a_s \cdot y_s$ and $a_{s+i} \cdot \hat{y}_{s+i}$, as well as the inputs fed to GRU cell in seq2seq model.

3 Experiments

In this section, we evaluate the performance of the proposed RMA-RNN algorithm for human motion prediction as well as the roles of its different modules.

3.1 Experimental Settings

In experiments, we followed previous works [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017], and focusd on the Human 3.6M dataset [Ionescu *et al.*, 2014], which is currently the largest human motion dataset for 3D mocap data analysis. Human 3.6M dataset provides 15 activities performed

	Walking					Eating					Smoking					Discussion				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ERD	0.93	1.18	1.59	1.78	2.24	1.27	1.45	1.66	1.80	2.02	1.66	1.95	2.35	2.42	3.61	2.27	2.47	2.68	2.76	3.16
LSTM-3LR	0.77	1.00	1.29	1.47	1.84	0.89	1.09	1.35	1.46	1.97	1.34	1.65	2.04	2.16	3.61	1.88	2.12	2.25	2.23	2.45
SRNN	0.81	0.94	1.16	1.30	1.78	0.97	1.14	1.35	1.46	2.09	1.45	1.68	1.94	2.08	2.64	1.22	1.49	1.83	1.93	2.24
DAE-LSTM	1.00	1.11	1.39	N/A	1.39	1.31	1.49	1.86	N/A	2.01	0.92	1.03	1.15	N/A	1.77	1.11	1.20	1.38	N/A	1.73
Zero-velocity	0.39	0.68	0.99	1.15	1.32	0.27	0.48	0.73	0.86	1.38	**0.26**	**0.48**	0.97	0.95	1.69	0.31	0.67	0.94	1.04	1.96
RRNN	0.33	0.55	0.70	0.77	0.98	0.26	0.43	0.66	0.80	1.38	0.35	0.64	1.11	1.16	1.90	0.35	0.74	1.04	1.09	1.77
RM (Ours)	0.31	0.50	0.67	0.74	0.95	0.24	0.36	0.57	0.72	1.18	0.27	0.49	**0.92**	**0.90**	**1.65**	0.31	0.65	**0.89**	**0.96**	1.86
RMA (Ours)	**0.28**	**0.45**	**0.62**	**0.68**	**0.79**	**0.21**	**0.34**	**0.53**	**0.68**	**1.16**	**0.26**	0.50	0.96	0.93	1.71	**0.29**	**0.64**	0.90	**0.96**	**1.72**
	Directions					Greeting					Phoning					Posing				
Zero-velocity	**0.39**	**0.59**	0.79	0.89	1.50	0.54	0.89	1.30	1.49	1.80	0.64	1.21	1.65	1.83	2.04	0.28	0.57	**1.13**	**1.37**	2.78
RRNN	0.44	0.70	0.84	0.94	1.46	0.55	0.90	1.33	1.51	1.93	0.61	1.15	1.53	1.66	**1.73**	0.45	0.83	1.50	1.74	**2.40**
RM (Ours)	0.42	0.65	0.81	0.90	1.45	0.55	0.88	1.30	1.47	1.82	**0.58**	1.14	1.50	1.64	1.81	0.29	0.60	1.16	1.40	2.59
RMA (Ours)	0.40	0.61	**0.77**	**0.86**	**1.42**	**0.52**	**0.86**	**1.26**	**1.43**	**1.79**	0.59	**1.11**	**1.47**	**1.59**	**1.73**	**0.26**	**0.54**	1.14	1.41	2.43
	Purchases					Sitting					Sitting down					Taking photo				
Zero-velocity	0.62	0.88	1.19	1.27	2.45	**0.40**	0.63	1.02	**1.18**	**1.63**	**0.39**	**0.74**	**1.07**	**1.19**	**1.90**	0.25	0.51	0.79	0.92	1.27
RRNN	**0.59**	**0.84**	1.25	1.32	2.38	0.48	0.79	1.27	1.48	2.12	0.52	0.98	1.52	1.74	2.57	0.33	0.63	0.98	1.11	1.51
RM (Ours)	0.64	0.86	**1.14**	1.21	2.37	0.43	**0.62**	**1.01**	**1.18**	1.68	0.43	0.77	1.10	1.22	**1.90**	**0.24**	**0.50**	**0.77**	**0.90**	**1.21**
RMA (Ours)	**0.59**	**0.84**	**1.14**	**1.19**	**2.33**	**0.40**	0.64	1.04	1.22	1.71	0.41	0.77	1.14	1.29	2.07	0.27	0.52	0.80	0.92	**1.21**
	Waiting					Walking Dog					Walking together					Average				
Zero-velocity	0.34	0.67	1.22	1.47	2.63	0.60	0.98	1.36	1.50	1.96	0.33	0.66	0.94	0.99	1.52	0.40	0.71	1.07	1.21	1.85
RRNN	0.37	0.70	1.16	1.34	2.33	0.54	0.89	1.22	1.35	**1.88**	0.30	0.60	0.85	0.88	1.38	0.43	0.76	1.13	1.26	1.85
RM (Ours)	0.34	0.67	1.16	1.37	2.46	0.55	0.89	1.24	1.38	1.91	0.29	0.56	0.77	0.81	1.38	0.39	0.67	1.00	1.12	1.75
RMA (Ours)	**0.33**	**0.65**	**1.12**	**1.30**	**2.28**	**0.53**	**0.87**	**1.16**	**1.33**	2.00	**0.28**	**0.52**	**0.68**	**0.71**	**1.31**	**0.37**	**0.66**	**0.98**	**1.10**	**1.71**

Table 1: Detailed results for human motion predictions on 15 actions from Human3.6M dataset in terms of mean Euler angle error on 80, 160, 320, 400ms (short-term) and 1000ms (long-term). Top section corresponds to previous RNN-based models. "Zero-velocity" is a baseline that constantly predicts last observed frame. "RM" stands for retrospection module, "RMA" stands for retrospection module with attention. The best result in bold. Our model outperforms other baselines in most scenarios.

by seven actors, including both periodic and aperiodic activities. Each activity trial consists of 3,000 to 5,000 frames. For each frame, 32 joints with a global translation and rotation are provided to represent the 3D human pose and each rotation is represented with its exponential map. We followed the standard data pre-processing for mocap data [Fragkiadaki *et al.*, 2015; Jain *et al.*, 2015; Li *et al.*, 2018; Martinez *et al.*, 2017]. Each pose would be normalized and global translation and rotation are set to zero. Joint angles dimensions that have constant standard deviation have been discarded to decrease computations, as they do not contribute to human dynamics. The final dimension of our input data is thus 54. We also down-sampled the original data by 2, making its sampling rate 25fps. The hyper parameter α in Eq. 7 is set to 0.5. Different from previous works that take activity labels as supervision information in the format of one-hot encoding [Martinez *et al.*, 2017], the proposed algorithm is an unsupervised method to model human dynamics.

Similar to previous works, we tested on subject 5 while the rest six subjects were used for training. During the training, we fed the network 50 frames (2 seconds in total), and predicted the subsequent 25 frames (1 second in total). We also trained a general model, where the input seed sequences are randomly selected from all the activities. Although error is minimized over 1 second, we let the network predict 2 seconds for qualitative comparison since RMA is able to capture long-term dependencies. Our RNN architecture was designed according to the suggestions in RRNN [Martinez *et al.*, 2017]. We adopted a single gated recurrent unit with 1024 units. Momentum method was used to optimize the proposed algorithm and the learning rate is set to 0.005. The batch size is set to 16, and gradient clipping to maximum L2-norm of 5. Our network was implemented using TensorFlow, and it takes 92ms per step on an NVIDIA Titan GPU.

3.2 Evaluation on Human3.6M Dataset

For a fair comparison, we evaluated the performance using the mean angle error for the 15 actions on subject 5 in Human 3.6M dataset and reported the error at 80ms, 160ms, 320ms and 400ms for short-term prediction as in [Martinez *et al.*, 2017], as well as 1000ms for long-term prediction as in [Li *et al.*, 2018]. Following [Martinez *et al.*, 2017], we also visualized the generated poses for qualitative analysis. State-of-the-art deep RNNs based approaches are included in comparison experiments, including ERD and LSTM-3LR [Fragkiadaki *et al.*, 2015], SRNN [Jain *et al.*, 2015], DAE-LSTM [Ghosh *et al.*, 2017] and RRNN and zero-velocity [Martinez *et al.*, 2017]. We used the official implementation to re-produce results of RRNN, and slightly better results were achieved than those reported in [Li *et al.*, 2018].

Quantitative Comparison

Table 1 shows our quantitative comparison with a set of baselines of human pose generation on 15 actions from Human 3.6m Dataset. Compared with ERD, LSTM-3LR, SRNN and DAE-LSTM, our model outperforms them in all the scenarios. To evaluate the effect of our retrospection module and spatial attention module, we mainly focused on comparisons with RRNN method and a strong zero-velocity baseline.

Compared with RRNN, our model outperforms it in almost all the scenarios. We can see that our retrospection module assists to produce more precise human poses in most cases, especially for long-term prediction (1000ms), comparing the second and third rows from the bottom (RRNN and RM), which highlights of difficulties for RNN to maintain complicated correlations on longer horizon. Thus, our retrospection

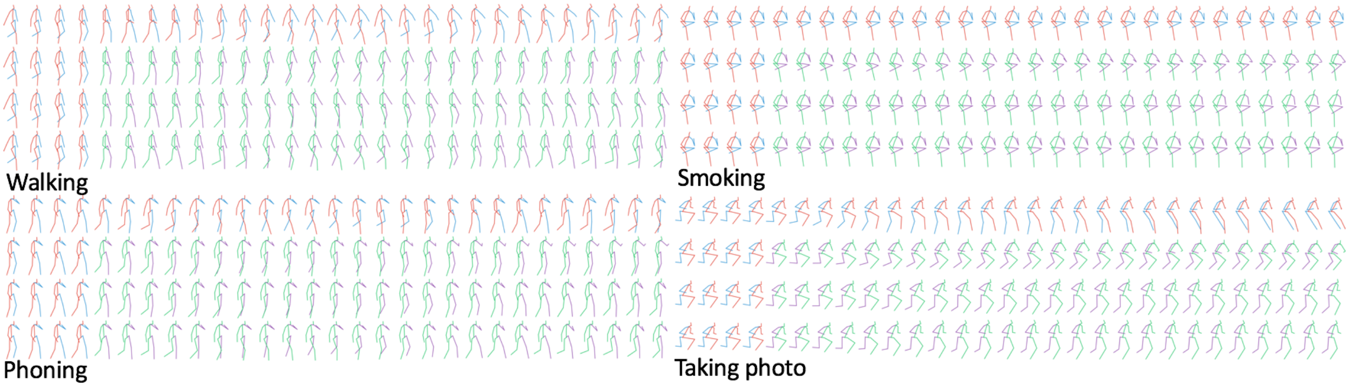


Figure 2: Qualitative motion generations for 2 seconds on different actions from Human3.6M dataset. The top sequence corresponds to ground truth, the second one to residual RNN, the third one to retrospection module and the bottom one to retrospection module with attention. The first four frames are the last four frames of observed seed sequence fed to the network. Our final model, retrospection module with attention on RNN (RMA-RNN), produces more realistic, continuous human motion predictions. Best viewed in color with zoom.

module enables it to keep track of information from distant past by keeping recalling previous subsequences. Meanwhile, our spatial attention module assists to further explore these complicated spatial dynamics and correlations by paying different attention on joints, comparing the last two rows (RM and RMA). With the combination of them, our model can produce more precise predictions (see the “Average” section).

Compared with zero-velocity baseline, our model still outperforms it in most scenarios. Although zero-velocity baseline generate static predictions, it outperforms most existing RNN-based methods especially in short-term prediction and aperiodic actions. For periodic actions(e.g. walking), our model consistently outperform it. For aperiodic actions which involve small movements in upper-body(e.g. smoking, purchases), our model still outperform it. For these highly aperiodic actions with high acceleration(e.g. posing, sitting down), it’s difficult for existing deep learning methods to capture its dynamics, however, our model still gain better quantitative results than other baselines.

Qualitative Comparison

Figure 2 shows our qualitative comparison with RRNN and ground truth on 4 actions. The sequences from top to bottom correspond to ground truth, RRNN, RM and RMA. Since our model can learn long-term dependencies during training, we expect the model to make predictions over longer horizons. Thus, though the model is trained to minimize the error over 1 second, we visualize 2 seconds prediction for 4 different types of actions to evaluate the performance of our model.

For periodic action(e.g. walking), both RRNN and our model generate predictions close to ground truth for the first second because the correlations of joints can be easily captured due to its periodic, mild dynamics. However, our RMA model can further explore the key joints which have more dynamics. With more attention to legs, our model produces continuous, dynamic predictions, whereas the others coverage to mean pose during the last several frames, which suggests that spatial attention module assists to further explore spatial correlations and generates human poses close to ground truth.

For aperiodic action with small movements(e.g. smoking), RRNN generates implausible human poses where both arms and legs move in the wrong directions after several frames, however, our models RM and RMA can both generate realistic poses which maintain the smoking gesture as the ground truth. Furthermore, for a combined action phoning where upper body involves small movements while lower body involve periodic movements, RRNN converges to mean pose and cannot maintain phoning gesture after 1 second, whereas RM and RMA continue generating realistic human poses.

For highly aperiodic action with complicated dynamics(e.g. taking photo), we visualize a challenging subsequence where the actor rises while holding the camera. RRNN quickly converges to mean pose where it sticks to squat gesture. RM and RMA, on the other hand, maintain the similar velocity with ground truth. However, for the next second, where the ground truth is transferred to walking, all the models fail to predict it due to its high uncertainty.

Finally, we compare RM with RMA to evaluate the performance of spatial attention module. For all the actions we visualize, we can see that our RMA-RNN model produces more realistic human poses while maintaining the similar velocity with ground truth, especially for long-term prediction, which suggests the combination of retrospection module and spatial attention modules generates better predictions.

User Studies

Following [Gui *et al.*, 2018], we generate 90 short term and long term motion sequences and ask judges to do pair-wise evaluation. Results are given in Table 2. We can outperform RRNN and be comparable with ground truth.

	Short-term			Long-term		
Pair	Ours	GT	RRNN	Ours	GT	RRNN
Ours	n/a	47.6%	75.6%	n/a	42.9%	95.1%
GT	52.4%	n/a	76.2%	57.1%	n/a	97.6%

Table 2: Each number represents the percentage that our generated sequence or ground truth is selected from a pair comparison.

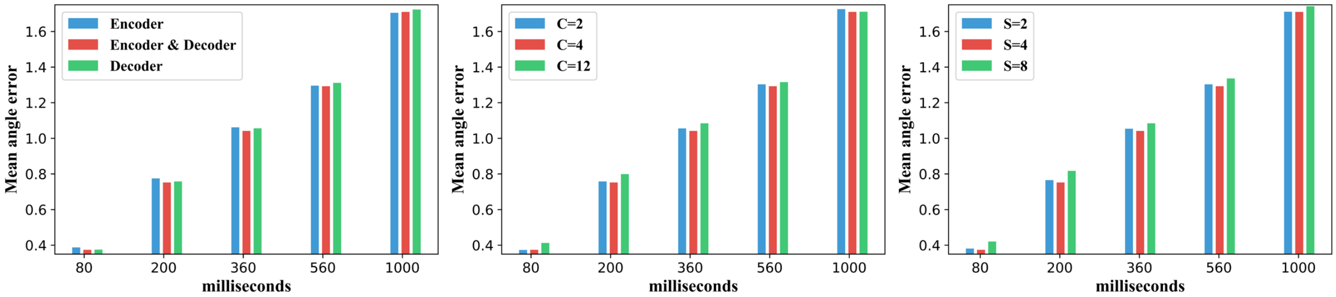


Figure 3: Ablation study for retrospection module. The left chart corresponds to the quantitative comparison of different location selections for anchor points and the one which sets anchor points on both encoder and decoder networks achieves the best performance. The middle chart corresponds to different interval sizes of anchor points, which shows both small and large interval sizes increase the error. The right chart corresponds to different size of subsequence to retrospect for each anchor point, which shows similar results with interval sizes.

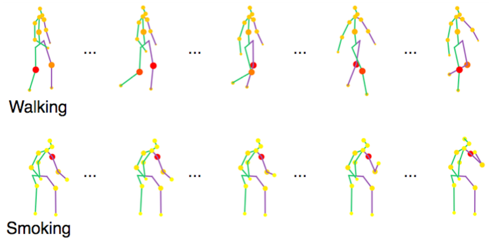


Figure 4: Visualization of spatial attention responses using heat map. For each joint, the size and color of circle represents the learned degree of importance, where small size with yellow indicates the low attention whereas large size with red indicates the high attention.

3.3 Ablation Study

The Role of Spatial Attention

We propose a spatial attention module in order to assist our model to learn spatial dependencies by paying more attention to key joints. To evaluate the performance of it, we apply heat map to the input sequence and visualize the key joints it finds. Figure 4 visualizes the effectiveness of it. For periodic action walking, more attention is paid to legs which contribute most movement. In addition, both legs receive more attention alternately, which indicates the spatial attention captures the periodic propriety of walking. For aperiodic action smoking, although it's nearly a static sequence except for small movement on one arm, spatial attention captures it successfully. As a result, with learned spatial attention on key joints, the model can generate more precise human poses, and with variation of attention distributions over time, spatial attention captures the tendency of human motion so that the model can further explore spatio-temporal correlations for different actions.

Ablation Study for Retrospection Module

In our retrospection module, we set anchor points over the whole sequence including encoder and decoder to retrospect previous frames. To further explore the effectiveness of anchor points on different locations, we design 3 scenarios: (a) anchor points are set only on encoder (b) set only on decoder (c) set on both of them. The results are shown in the left chart in Figure 3. Results suggest that the anchor points on encoder network contribute more to minimize the errors in long-term prediction. Meanwhile, those on decoder network contribute more to short-term prediction. With anchor points set on both encoder and decoder, the model achieves complementary quantitative results and better performance.

In addition, we set anchor points on RNN every fixed number of frames C to show the influence of interval size of anchor points, shown in the middle chart in Figure 3. We further explore the influence of the size of subsequence S to retrospect for each anchor point, shown in the right chart of Figure 3. We find $C = 4$ is the best choice to minimize the errors over the whole horizon. Smaller size S indicates RM module retrospects only a part of the subsequence and may lose some key information. However, larger length to look back results in redundant information, which makes it difficult to learn spatio-temporal correlations. Thus, with retrospection size corresponding to interval size, retrospection module exactly covers the entire sequence and gains the best performance.

4 Conclusions

In this paper, we propose a retrospection module with attention to address the challenges in human motion prediction. We demonstrate that the current RNN-based approach cannot produce plausible human poses for aperiodic actions and converges to mean pose quickly, whereas our model eliminates these limitations and outperforms it by setting anchor points to retrospect previous frames and applying spatial attention upon joints. Based on quantitative and qualitative evaluations, we show the effectiveness of retrospection module and spatial attention module, which together capture complicated spatio-temporal correlations, invariant and dynamical information. Our proposed model RMA-RNN, focusing on learning long-term dependencies, can be trained on large mocap datasets in unsupervised manner with less parameter tuning and generates longer, more realistic and coherent predictions.

Acknowledgements

This work was supported in part by the Australian Research Council under Project DE180101438 and FEIT Early Career Researcher Development Grant.

References

[Akhter *et al.*, 2012] Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. Bilinear spatiotemporal basis models. *ACM Trans. Graph.*, 31(2):17:1–17:12, April 2012.

[Barsoum *et al.*, 2017] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. *CoRR*, abs/1711.09561, 2017.

[Chen *et al.*, 2018] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018.

[Cho *et al.*, 2014] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014.

[Fragkiadaki *et al.*, 2015] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4346–4354, Washington, DC, USA, 2015. IEEE Computer Society.

[Gehring *et al.*, 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122, 2017.

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

[Gui *et al.*, 2018] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *European Conference on Computer Vision (ECCV). Munich, Germany (September 2018)*, 2018.

[Ionescu *et al.*, 2014] C. Ionescu, D. Papava, V. Olaru, and C. 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, July 2014.

[Jain *et al.*, 2015] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. *CoRR*, abs/1511.05298, 2015.

[Lehrmann *et al.*, 2014] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014.

[Li *et al.*, 2018] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. *CoRR*, abs/1805.00655, 2018.

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

[Nie *et al.*, 2018] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018.

[Pavlovic *et al.*, 2000] Vladimir Pavlovic, James M. Rehg, and John MacCormick. Learning switching linear models of human motion. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 942–948, Cambridge, MA, USA, 2000. MIT Press.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[Tekin *et al.*, 2016] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Fusing 2d uncertainty and 3d cues for monocular body pose estimation. *CoRR*, abs/1611.05708, 2016.

[Urtasun *et al.*, 2008] Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D. Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1080–1087, New York, NY, USA, 2008. ACM.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, Yangxi Li, and Chao Xu. Large-margin multi-view gaussian process for image classification. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 7–12. ACM, 2013.