

Received July 13, 2019, accepted July 22, 2019, date of publication August 5, 2019, date of current version August 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933221

Spatial–Temporal Attention-Based Human Dynamics Retrospection

MINJING DONG AND CHANG XU^{ID}

Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia

Corresponding author: Chang Xu (c.xu@sydney.edu.au)

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

ABSTRACT Motivated by impressive success of deep recurrent neural networks (RNNs), sequence-to-sequence (seq2seq) architecture has been widely adapted to tackle human motion prediction. However, forecasting in longer time horizons always leads to implausible human poses or converges to mean poses. To address these challenges, we dig into the root causes and lay emphasis on two key principles. First, error can be easily accumulated on seq2seq architecture without modifications and thus RNNs cannot recover from its own mistakes in longer time horizons. Second, all the frames or joints are treated equally, whereas both of them often have different levels of importance in human motion. To mitigate this gap, we propose to retrospect human dynamics with attention. We design a retrospection module built upon seq2seq architecture to recollect previous subsequences and correct mistakes in time which enables a self-correction ability. This assists the original seq2seq architecture to eliminate error accumulation which improves significantly both short-term and long-term performances. Besides, we present two attention techniques to explore correlations among different joints as well as different frames in both spatial and temporal domains, which successfully capture key properties of different actions and enable our model to generate more realistic human poses. Quantitative and qualitative experiments have been both conducted to evaluate the our proposed model. Experimental results clearly demonstrate the superiority of proposed model over other baselines.

INDEX TERMS Human motion prediction, predictive models, recurrent neural networks, attention mechanisms.

I. 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 and decision making. 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 [1], [2], large-scale human motion datasets

are available for investigating machine learning approaches in human dynamics analysis, such as Human 3.6M dataset [3] and CMU dataset. In this work, we mainly focus on skeleton-based mocap datasets where human poses are represented as joint rotations.

A wide variety of RNN-based methods have emerged to tackle the human motion analysis problem [4]–[7] due to the success of Recurrent Neural Networks (RNNs). For example, Encoder-Recurrent-Decoder network has been proposed to learn temporal dependencies, with spatial encoder and decoder wrapped upon recurrent cell and last hidden state encoding human poses to generate motion prediction [4]. Besides directly encoding human poses by hidden variables, Residual RNN models velocity representations, which boosts the short-term prediction performance and solves the first frame discontinuity by applying sequence-to-sequence model with residual connections [7].

Though these methods have achieved impressive performance in analyzing human motion, there are still some drawbacks. Firstly from the temporal aspect, these RNN-based

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

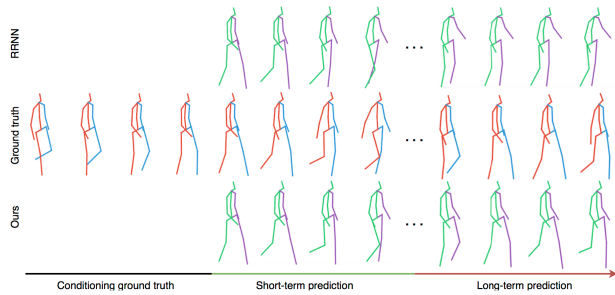


FIGURE 1. Seed motion sequence is fed to the network, and shown in the left frames. The right frames correspond to RRNN, ground truth and the proposed algorithm from top to bottom in both short-term and long-term predictions. RRNN converges to mean pose whereas our model still sticks to ground truth over longer time horizon.

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 and thus converges to mean pose. In addition, some actions might require more information from distant past instead of recent one, however, the current prediction can be easily overwhelmed by recent frames if we simply decode the last hidden representation. On the other hand, from the spatial aspect, 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 [8], and they often get stuck with the mean pose problem, as shown in Figure 1. It is obvious that residual RNN [7] loses its velocity quickly and make almost static prediction over longer horizon.

To address these challenges, we propose to retrospect human dynamics with attention. To enable self-audit, a retrospection module is designed to recall previous subsequences and make timely adjustments based on them so that the error will be corrected instead of accumulated. Furthermore, to produce plausible human poses, we introduce two attention mechanisms, one focusing on key joints during human motion to learn spatial dependencies and the other pay more attention to key frames and summarizes historical information from encoder. With the assistance of these modules, our proposed algorithm successfully mitigates the gap of performance between seq2seq architecture and ground truth. As shown in Figure 1, our model can make satisfactory prediction with similar velocity as ground truth.

This work is an expanded version of conference publication [9]. Compared with previous version of this work, we make following extensions: 1) We enrich the content of abstract, introduction and related work to cover sufficient details. 2) We propose a new temporal attention module to explore the temporal correlations of human motion, which cooperates with spatial attention proposed in previous version

to further improve the performance. 3) Besides Human 3.6M dataset, We conduct more experiments on CMU dataset which is widely used for human motion analysis. We provide more experiment results to further demonstrate the superiority of our model over other baselines and its generalization ability. 4) We provide more details on ablation studies for spatial attention and visualize heat map on more actions to illustrate the effectiveness of it. 5) We add ablation study for temporal attention by observing the temporal attention score variations to demonstrate that temporal attention is able to capture properties of different actions. 6) We add more experiments under different settings to understand the impact of different modules and their contributions. 7) We add more experiments to discuss the impact of hyper-parameter and provide appropriate value range.

Our main contributions in this work can be summarized as:

First, we introduce a retrospection module which takes care of previous mistakes and makes correction. By setting anchor points upon the whole sequence, retrospection module not only enables a timely self-correction, but also strengthens the confidence in making the subsequent prediction. It is necessary for the model to look back in order to move forward because seq2seq architecture cannot recover from accumulated error in decoder and gets stuck in local optimum without self-audit. With retrospection module, our model have strong confidence in generating more realistic human dynamics.

Second, we design a spatial attention module to distinguish importance of different joints in each frame so that more attention can be paid on joints that involve more movements. Thus, some complicated motion properties can be further exploited for human dynamics, such as motion periodicity and movement tendency, which enables our model to capture these spatial dependencies.

Third, a temporal attention module is designed to explore the degrees of importance of different frames in encoder. Due to the high uncertainty, the prediction might requires more information in distant past. Instead of directly send decoder the last hidden representation, we assign importance scores on hidden representations on different time steps to generate a context which summaries all historical information so that our model can learn temporal dependencies easily to further generate more realistic human motions.

Finally, we perform both quantitative and qualitative experiments showing that our proposed algorithm retrospection module with attention on RNN (RMA-RNN) outperforms RRNN and other RNN-based approaches. Equipped with RMA, seq2seq architecture produces more realistic and coherent human motion predictions, especially in longer time horizon.

II. RELATED WORK

Different approaches have been proposed for human motion prediction, including traditional methods and deep learning based methods. In this section, we give a overview of related literature.

A. MODELING OF HUMAN MOTIONS

Uncertainty of human motions and complicated correlations make learning human dynamics and generation of naturalistic human poses difficult to accomplish. Traditional approaches focus on hidden Markov models [10], linear dynamical systems [11], Gaussian process latent variable models [12], [13] and bilinear spatio-temporal basis models [14]. However, there are some limitations. For example, Restricted Boltzmann Machine models [15]–[18] require complicated training process due to learning probabilistic models and also require sampling for approximate inference. Thus, there exists trade-offs between model capacity and inference complexity, which make them difficult to be trained on large datasets.

B. DEEP LEARNING METHODS FOR HUMAN DYNAMICS

Recurrent neural networks show great success in different areas, such as machine translation [19], speech recognition [20], sequence generation [21], and image caption [22]. Motivated by advances of them, RNNs have been widely used on human motion prediction. Fragkiadaki *et al.* [4] propose a 3-layer of Long Short-Term Memory network (LSTM-3LR) and an Encoder-Recurrent-Decoder network (ERD) which adds non-linear, non-recurrent spatial encoder and decoder to recurrent cell so that they jointly learn a representation of human pose and capture temporal correlations. Jain

hboxetal [6] propose Structural-RNN model (SRNN) with a manual spatio-temporal graph which encodes semantics of human motion as input to RNNs. Both of them are action-specific models and add noise scheduling to handle error accumulation which means that RNN is not able to recover from its own mistakes [23], however, noise scheduling relies heavily on hyper-parameter tuning and hurts short-term prediction due to discontinuity while action specific models become inefficient for large mocap datasets because different models need to be trained for different actions.

Recently, some general models with less hyper-parameter tuning have been proposed. Ghosh *et al.* [5] introduce Dropout Autoencoder LSTM (DAE-LSTM) which leverages de-noising autoencoders with dropout to learn spatial dependencies, Martinez *et al.* [7] introduce sequence-to-sequence model with sampling-based loss and simply add a residual connection (RRNN) in order to model velocity representations, which solves the first frame discontinuity and requires no parameter tuning, but the predictions converge to mean pose quickly. To deal with such a limitation, GAN networks are also proposed, Barsoum *et al.* [24] are 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. Gui *et al.* [25] introduce fidelity and continuity discriminators with geodesic loss to boost the performance, however, it adds one-hot vector to the input indicating the action class, which makes it supervised method and limits the usage of it in different scenarios.

Besides RNN-based methods, fully-connected networks and convolutional neural networks (CNN) are also used to capture dynamic information of human motion. Büttepage *et al.* [26] develop an unsupervised deep representation learning using fully-connected networks with a bottleneck encoding-decoding structure. Li *et al.* [3] propose a convolutional sequence-to-sequence model with short-term and long-term encoder to capture both invariant and dynamical information. With comparison to RRNN, it gains better results but still hard to outperforms zero-velocity baseline in most scenarios. To this end, our goal is to develop a general RNN-based model with less hyper-parameter tuning in unsupervised manner by retrospecting human dynamics with spatial attention, which generate more precise human poses in both short-term and long-term predictions.

C. ATTENTION MECHANISM

Attention mechanisms have been widely applied in neural machine translation [19], [27]. Driven by the great performance of it, attention mechanisms are involved in various human motion analysis tasks recently. In terms of sequence generation problems, temporal attention has been applied to human dynamics. Tang *et al.* [28] propose a modified highway unit (MHU) which filters motionless joints by summarizing historical sequence using temporal attention mechanism, but the performance is still unsatisfactory. A spatial soft attention has been introduced to solve image caption [29], which we find fit human motion prediction. With spatial soft attention on different joints, the model can learn different degrees of importance for each joint on different time steps. Furthermore, with temporal attention on different observed frames, the model can easily capture temporal dependencies of human motions.

III. METHODOLOGY

We adapt sequence-to-sequence (seq2seq) architecture [30], which is widely used in recent RNN based methods for motion sequence generation, as shown in Figure 2. It consists of two networks, an encoder 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.

A. PROBLEM FORMULATION

Formally consider an observed seed sequence of human poses $X_{1:t} = [x_1, x_2, \dots, x_t]$, 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 GRU cell 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

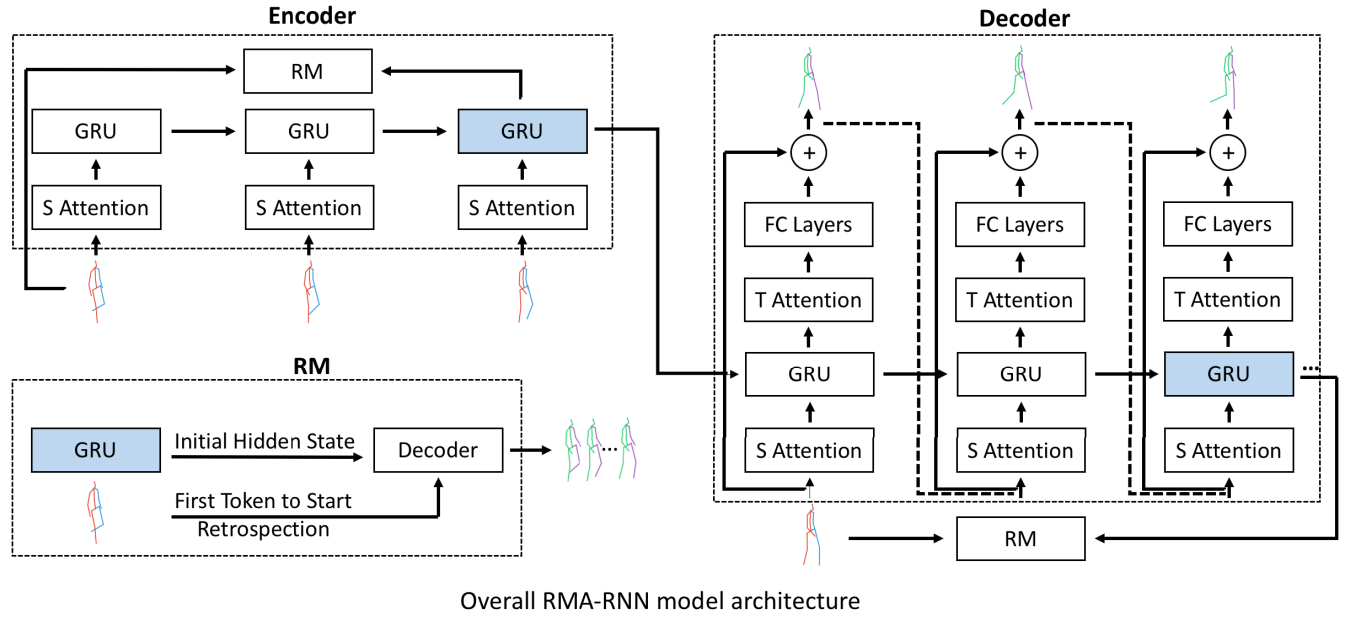


FIGURE 2. An illustration of our retrospection module with attention on RNN model (RMA-RNN). The red-blue skeletons represent the ground truth, and the green-purple skeletons represent the prediction. GRU cells with blue background are anchor points selected, which are initial states of retrospection module (RM). We introduce spatial attention to focus on joints with more movements and temporal attention to focus on frames with more important information in Fig. 3. 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.

sequence as

$$\mathcal{L}_{seq} = \frac{1}{T} \sum_{t'=t+1}^{t+T} \|\hat{x}_{t'} - x_{t'}\|_2^2. \quad (1)$$

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 [3]. To handle error accumulation and capture spatial dynamics accurately, we propose a retrospection module with attention upon chain-structured GRU. Residual connection [7] 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 2.

B. 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

$$P_{1:n} = h_{1:C} = [h_C, h_{2C}, \dots, h_{nC}], \quad (2)$$

where n is the number of anchor points, and nC is less than $t + T$. Figure 2 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 2(a). For expression simplicity, we adopt a new variable y to represent elements in the sequence $\{x_1, \dots, x_t, \hat{x}_{t+1}, \dots, \hat{x}_{t+T}\}$. Given the k -th anchor point, we select the subsequence $Y = \{y_{(k-1)C+1}, \dots, y_{kC}\}$, where the length of the subsequence has been fixed as C . The hidden state is calculated as,

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

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

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

where function f represents two fully-connected layers forward operation. According to [7], residual connection is also adopted in Eq. (4), so that $f(\hat{h}_{s+1})$ can represent the velocity

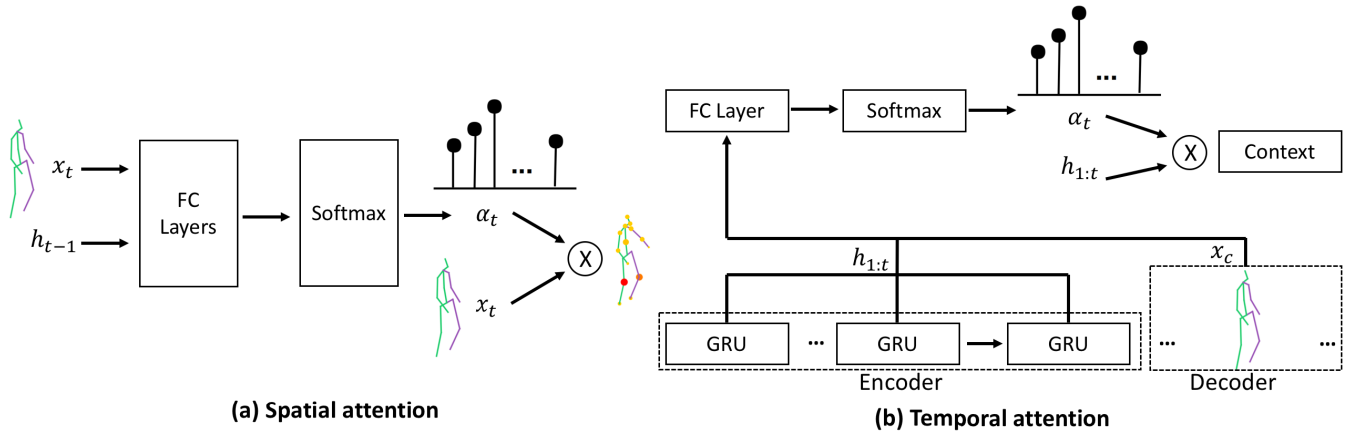


FIGURE 3. An illustration of our proposed attention modules including spatial attention and temporal attention. The current input and last hidden representation are fed to fully-connected layers with Softmax layer to compute spatial attention weights (Fig. (a)). With the current input and all the hidden states from encoder, temporal attention module computes temporal attention weights and generates context which summarizes all the historical information (Fig. (b)).

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 $\{\hat{y}_{s+1}, \dots, \hat{y}_{s+C-1}\}$, the predicted loss for the k -th retrospect module can be computed as

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

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

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

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.

C. SPATIAL ATTENTION

We introduce spatial attention module to explore different importance of joints in frames. Spatial attention weights are assigned 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, which contains more information to model this activity, “smoking”, however, involves more joint angle rotation of arms. Spatial correlations of different joints can therefore be explored and exploited to generate more plausible human poses.

Figure 3(a) 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 $x_t = [x_{t,1}, x_{t,2}, \dots, x_{t,K}]$, the spatial 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 weakened 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.

D. TEMPORAL ATTENTION

Besides spatial attention applied on each joint, we introduce temporal attention module to summarize all the historical information from different observed frames. Temporal attention weights are assigned to all the frames on encoder so that frames which contain more information to generate current frame will received more attention and thus assists our model to capture temporal dependencies.

Figure 3(b) illustrates the framework of the proposed temporal attention module. Consider the GRU's hidden states on encoder denoted as $h[1 : t]$ and the current input denoted as x_c , the temporal attention scores at time step c can be computed as

$$\text{score}(c, i) = W_s[x_c, h_i] + b_s, \quad (10)$$

where $i \in [1, t]$, W_s is weight matrix and b_s is the bias vector. Similarly, the scores are normalized by a Softmax layer as

$$a_{c,i} = \frac{\exp(\text{score}(c, i))}{\sum_{m=1}^t \exp(\text{score}(c, m))}, \quad (11)$$

With attention scores on $h[1 : t]$, we can obtain a context at time step c which summarizes all the historical information in encoder with attention on different frames.

$$\text{context}(c) = \sum_{i=1}^t a_{c,i} \cdot h_i, \quad (12)$$

To generate the next frame, instead of directly feeding the decoder with hidden state, a combination of context and current hidden state is fed to linear layers to model human dynamics. Thus, the inputs fed to linear layers in Eq. (4) P_k and $\hat{h}_{s+1}$ will be replaced by $[\text{context}(s), h_s]$ and $[\text{context}(s+1), h_{s+1}]$, as well as the inputs fed to decoder in seq2seq model.

IV. 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.

A. EXPERIMENTAL SETTINGS

1) DATASET AND DATA PRE-PROCESSING

In experiments, we followed previous works [4], [6], [7], and focusd on the Human 3.6M dataset [31], which is currently the largest human motion dataset for 3D mocap data analysis. Human 3.6M dataset provides 15 activities performed by seven actors, including periodic activities such as “walking” as well as aperiodic activities such as “posing”. 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, as an exponential map.

We followed the standard data pre-processing for mocap data [3], [4], [6], [7]. Each pose would be normalized. Joint angles dimensions that have constant standard deviation have been discarded to decrease computations, as they do not contribute to human dynamics. In addition, global translation

and rotation are set to zero, since they are not considered in the error calculation. The final dimension of our input data is thus 54. We also down-sampled the original data by 2 as suggested, making its sampling rate 25fps. The hyper parameter α in Eq. III-B is set to 0.5. Different from previous works that take activity labels as supervision information in the format of one-hot encoding [7], the proposed algorithm is an unsupervised method to model human dynamics.

In order to evaluate the generalization performance of the proposed algorithm, we also tested it on the CMU Motion Capture dataset,¹ which has the similar data structure with the Human3.6M dataset [3].

2) BASELINES

State-of-the-art deep RNNs based approaches are included in comparison experiments, including ERD and LSTM-3LR [4], SRNN [6], DAE-LSTM [5] and RRNN and zero-velocity [7]. Since there is no public official implementations for most of these methods, we quoted results of ERD, LSTM-3LR and SRNN from [7] and results of DAE-LSTM from [5]. We used the official implementation to re-produce results of RRNN, and slightly better results were achieved than those reported in [3].

3) IMPLEMENTATION

Similar to previous works [3], [4], [6], [7], 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, instead of a model that corresponds to a particular activity. Therefore, we consider a more challenging task that training a multi-action model in an unsupervised manner to minimize the mean error of next 1 second prediction.

Our RNN architecture was designed according to the suggestions in residual RNN [7]. We adopted a single gated recurrent unit (GRU) [32] 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 16, and we also adopted 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.

B. 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 [7], as well as 1000ms for long-term prediction as in [3]. Table 1 shows our quantitative comparison with a set of baselines of human pose generation on actions walking, eating, smoking and discussion, which are commonly used for comparison in previous works. We then compare the proposed algorithm with

¹Dataset available at <http://mocap.cs.cmu.edu>

TABLE 1. Detailed results for human motion predictions on walking, eating, smoking and discussion 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, which produces static prediction. “RM” stands for retrospection module. “RMA” stands for retrospection module with attention. The best result in bold. Our model outperforms other baselines in almost all the scenarios.

	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 [4]	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 [4]	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 [6]	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 [5]	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 [7]	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 [7]	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)	0.26	0.42	0.60	0.66	0.79	0.21	0.34	0.52	0.66	1.17	0.27	0.49	0.90	0.89	1.62	0.29	0.63	0.89	0.95	1.72

TABLE 2. Detailed results for zero-velocity baseline, residual RNN and our models on the rest actions from Human3.6M dataset in terms of mean Euler angle error on 80, 160, 320, 400ms (short-term) and 1000ms (long-term). The best result in bold. Our model outperforms other baselines in most scenarios.

	Directions					Greeting					Phoning					Posing				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Zero-velocity [7]	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 [7]	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)	0.39	0.61	0.76	0.84	1.37	0.52	0.87	1.27	1.42	1.77	0.59	1.09	1.44	1.57	1.62	0.30	0.61	1.22	1.49	2.43
	Purchases					Sitting					Sitting down					Taking photo				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Zero-velocity [7]	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 [7]	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	0.83	1.12	1.19	2.29	0.40	0.65	1.05	1.23	1.74	0.42	0.80	1.18	1.33	2.07	0.26	0.53	0.81	0.94	1.22
	Waiting					Walking Dog					Walking together					Average				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Zero-velocity [7]	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 [7]	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)	0.32	0.63	1.10	1.29	2.37	0.53	0.86	1.17	1.34	1.91	0.26	0.52	0.70	0.72	1.24	0.37	0.66	0.98	1.10	1.69

RRNN method and zero-velocity baseline on the other 11 actions, as shown in Table 2. Compared with ERD [4], LSTM-3LR [4], SRNN [6] and DAE-LSTM [5], 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 state-of-the-art 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 long-term information with complicated correlations. Thus, our retrospection 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 RMA model can produce more precise predictions (see the “Average” section). Even compared with strong zero-velocity baseline which constantly predicts the last observed frame, our model still outperforms it in most scenarios, including periodic actions such as walking as well as some challenging actions such as discussion and purchases.

Following [3], [4], [6], [7], we also visualized the generated poses for qualitative analysis. Figure 4 shows our qualitative comparison with RRNN and ground truth on several actions. The sequences from top to bottom correspond to ground truth, RRNN, retrospection module and retrospection module with attention. Since our model can capture spatio-temporal correlations and 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, phoning), 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, RRNN quickly converges to mean pose and tend to make static prediction in walking and cannot maintains the pose (making phone call) the after one second, on the contrary, our proposed model RM and RMA can both stick to ground truth. Furthermore, with attention techniques, the human motion generated by RMA has more accurate velocity compared with RM.

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, on the contrary, our models RM and RMA can both generate

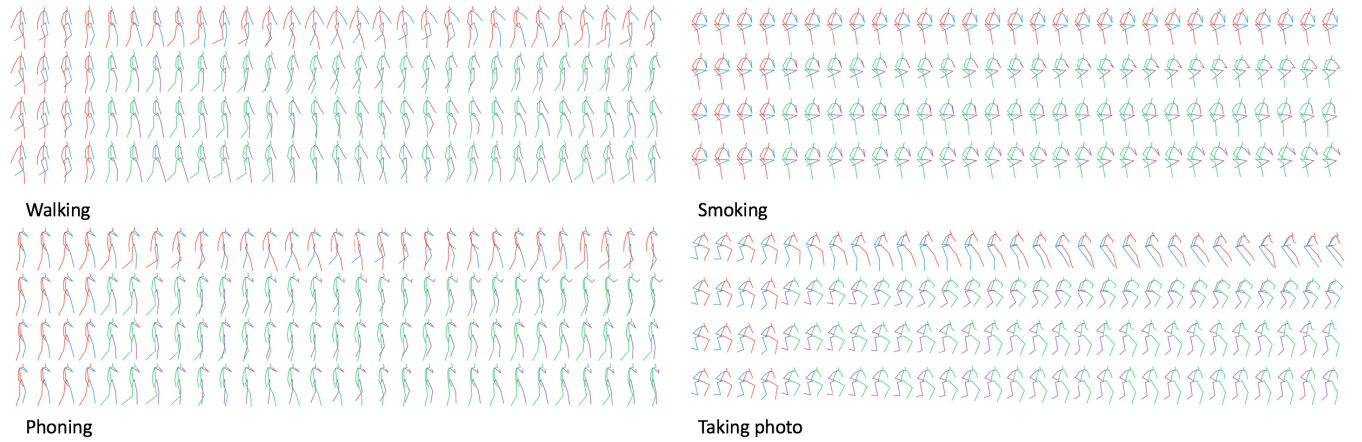


FIGURE 4. Qualitative motion generations for 2 seconds on different actions from Human3.6M dataset. The top sequence corresponds to ground truth, the second one to residual RNN, the third one to retrospection module and the bottom one to retrospection module with attention. The first four frames are the last four frames of observed seed sequence fed to the network. The RRNN produces unrealistic motions for smoking, phoning, and converges to mean poses for taking photo. Retrospection module yields more reasonable long-term predictions, comparing the second and third rows. Spatial attention module enables model to keep pace with ground truth, comparing the third and forth rows. Our final model, retrospection module with attention on RNN (RMA-RNN), produces more realistic, continuous human motion predictions. Best viewed in color with zoom.

TABLE 3. Detailed results for zero-velocity baseline, residual RNN and our model on 7 actions from CMU dataset in terms of mean Euler angle error on 80, 160, 320, 400ms (short-term) and 1000ms (long-term). The best result in bold. Our model outperforms other baselines in most scenarios.

	Walking					Running					Directing traffic					Soccer				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Zero-velocity [7]	0.41	0.60	0.83	0.95	1.34	0.56	0.99	1.38	1.46	1.54	0.30	0.57	0.90	1.01	2.19	0.27	0.48	0.92	1.10	1.84
RRNN [7]	0.36	0.49	0.58	0.64	0.89	0.34	0.50	0.64	0.68	0.72	0.35	0.65	1.07	1.21	1.91	0.28	0.47	0.83	0.95	1.59
RMA (Ours)	0.34	0.44	0.48	0.52	0.71	0.29	0.43	0.55	0.61	0.55	0.28	0.56	0.87	0.97	1.96	0.24	0.39	0.72	0.83	1.53
	Basketball					Jumping					Basketball signal					Average				
milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Zero-velocity [7]	0.49	0.82	1.40	1.64	2.83	0.36	0.62	1.48	1.68	2.03	0.24	0.44	0.75	0.86	1.17	0.38	0.65	1.10	1.24	1.85
RRNN [7]	0.47	0.78	1.26	1.42	2.12	0.58	1.05	1.93	2.11	2.21	0.42	0.78	1.32	1.51	1.96	0.40	0.68	1.09	1.22	1.63
RMA (Ours)	0.42	0.69	1.12	1.27	1.77	0.45	0.75	1.58	1.78	1.96	0.27	0.50	0.89	1.04	1.46	0.33	0.54	0.88	1.00	1.42

realistic poses which maintain the smoking gesture as the ground truth. In addition, 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. This suggests that retrospection module assists to improve the memory of RNN and eliminate error accumulation.

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.

C. EVALUATION ON CMU DATASET

Besides Human 3.6M Dataset, we also consider training our model on CMU dataset to evaluate the generalization ability of RMA model. We select 7 actions including periodic actions such as running as well as aperiodic actions such as basketball. Since both datasets have similar data structure, we pre-process the data and evaluate the quantitative results in the same way as in Section IV-A.1. For a fair comparison, we train RMA and RRNN under the same settings and

report the average angle error, shown in Table 3. Compared with RRNN, our model outperforms it in almost all the scenarios. Compared with zero-velocity baseline, our model still achieves better results in most scenarios. By applying RMA, the generated motion prediction tends to have a larger margin over longer horizons shown in “Average” column, which demonstrates the advantage of our proposed model in long term prediction.

D. ABLATION STUDY

1) ABLATION ANALYSIS ON DIFFERENT MODULES

In addition to the comparisons between retrospection module (RM) and retrospection module with attention (RMA) in Table 1 and 2, we further conduct thorough experiments to explore contributions of different modules to the final performance in Table 4. More importantly, combinations of different modules always obtain better quantitative results, which indicates that different modules can complement each other well. With the retrospection module equipped with spatial and temporal attention techniques, our full RMA model achieves the best performance.

2) 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

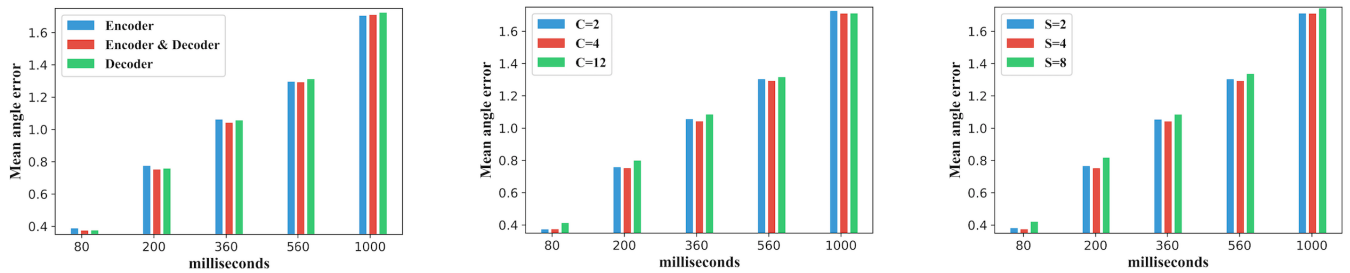


FIGURE 5. Ablation study for retrospection module. The left chart corresponds to the quantitative comparison of different location selections for anchor points and the one which sets anchor points on both encoder and decoder networks achieves the best performance. The middle chart corresponds to different 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.

TABLE 4. Thorough ablation analysis on retrospection module, spatial attention and temporal attention. Each row shows average angle error evaluated on Human 3.6M dataset under different module combinations.

Settings			Average				
RM	SA	TA	80	160	320	400	1000
			0.43	0.76	1.13	1.26	1.85
	✓		0.38	0.68	1.01	1.16	1.71
	✓	✓	0.38	0.67	0.99	1.12	1.71
✓			0.39	0.67	1.00	1.12	1.75
✓	✓		0.37	0.66	0.98	1.10	1.71
✓	✓	✓	0.37	0.66	0.98	1.10	1.69

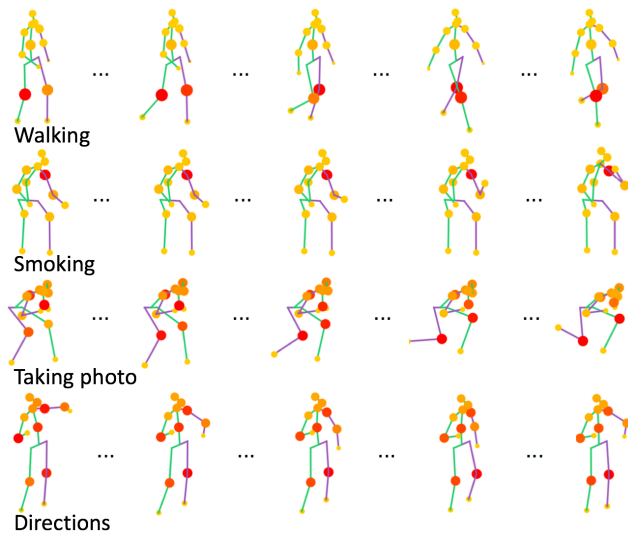


FIGURE 6. Visualization of spatial attention responses using heat map. For each joint, the size and color of circle represents the learned degree of importance, where small size with yellow indicates the low attention whereas large size with red indicates the high attention. Best viewed in color.

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 6 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 taking photo, we can see an attention migration from upper body

to lower body due to squatting movement. For action with small movement or low velocity, the attention responses distributions are similar for the whole input sequence but still pay more attention to key joints. For smoking, although it's nearly a static sequence except for small movement on one arm, spatial attention captures it successfully. And for directions, most limbs are involved in the movement and all of them receive high attention. 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.

3) THE ROLE OF TEMPORAL ATTENTION

The temporal attention module is designed to establish a direct connection between a specific frame in decoder and frames in encoder so that some properties of actions or some key frames can be captured to further assist our model to learn temporal dependencies by summarizing all the historical information in encoder. To illustrate the effectiveness of it, we record some temporal attention scores of frames on different locations from different actions and make them a line chart, as shown in Figure 7. For periodic action walking, we report the attention scores from the first frame (blue line) and the twentieth frame (orange line). There exist some periodic variations as the attention scores decrease and increase over time horizon. Furthermore, compared with the first frame and the twentieth frame, temporal attention scores tend to increase for 0–20 frames in encoder when the model make long term prediction, which denotes that temporal attention module assists our model to learn long-term dependencies. For aperiodic action smoking, some key frames will be paid more attention to, such as 15–20 frames in encoder. For directions which involve movement in almost all the frames, attention scores assigned to encoder frames have small variations.

4) THE ROLE OF ANCHOR POINTS

In our retrospection module, we set anchor points over the whole sequence including encoder and decoder to retro-

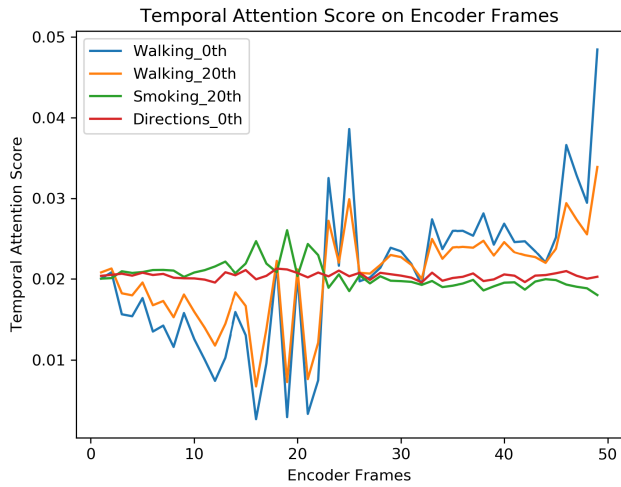


FIGURE 7. Ablation analysis on temporal attention. Each line represents the temporal attention scores on different encoder frames corresponding to a specific frame in decoder. Each label denotes its action type and the location of frame in decoder.

spect previous frames. To further explore the effectiveness of anchor points on different locations, we design 3 scenarios: (a). anchor points are set on encoder (b). anchor points are set on decoder (c). anchor points are set on whole sequence (both encoder and decoder). We compare the mean angle errors of them on different time steps, shown in the left chart in Figure 5. The 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 minimize the errors in short-term prediction. Thus, with anchor points on both encoder and decoder networks, the model achieves complementary quantitative results and better performance.

Furthermore, we find different interval sizes result in different performance. In the experiments, we set anchor points on RNN every fixed number of frames C . From the middle chart in Figure 5, we can see that the one with large interval size $C = 12$ has good performance in long-term prediction but much worse in short-term prediction, while the one with interval size of 2 has opposite performance, which is good in short-term but worse in long-term. With these observations, we argue that an appropriate interval size is needed for better performance, which can not be too small for model to learn long-term dependencies or too large to generate plausible short-term predictions. Thus $C = 4$ is the best choice to minimize the errors over the whole horizon.

Besides the interval size, we also consider the size of subsequence S to retrospect for each anchor point. In the right chart of Figure 5, we show the results for $S = 2$, $S = 4$ and $S = 8$. Recall that we have selected 4 as the interval size. We can see that both of $S = 2$ and $S = 8$ have worse performance than $S = 4$ which is consistent with interval size. Smaller size indicates RM module retrospects only a part of the subsequence and may lose some key information. However, larger size of subsequence to look back results

TABLE 5. Ablation studies on impact of $\mathcal{L}_{RM}$. Each row represents mean angle error under different α values.

α	80	160	320	400	1000
0.1	0.38	0.67	1.00	1.12	1.71
0.5	0.37	0.66	0.98	1.10	1.69

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.

5) THE ROLE OF RETROSPECTION MODULE

Our proposed RMA-RNN model contains few hyper-parameters. Besides C and S discussed in Section IV-D.4, we introduce α to balance the effect of RM and main seq2seq loss in Eq. III-B. In this section, we provide more experiment results on this hyper-parameter α . Generally, the impact of RM should not overwhelm the main seq2seq loss since retrospection module acts as an assistance of GRU to make it aware of its own mistake and enable a self-correction ability. In the experiments, appropriate α value ranges within $[0.1, 1.0]$. Impacts of α with different values on quantitative results are reported in Table 5.

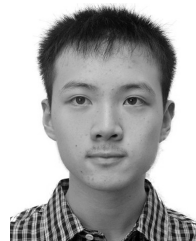
V. CONCLUSION

In this work, 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 in longer time horizons or 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 and temporal attention techniques upon joints and frames. Our proposed model can be treated as extended module built upon chain-structured RNN and can be easily applied on other RNN-based architectures. Based on quantitative and qualitative evaluations, we show the effectiveness of retrospection module and attention modules, 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.

REFERENCES

- [1] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2100–2108.
- [2] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2D and 3D image cues for monocular body pose estimation," Nov. 2016, *arXiv:1611.05708*. [Online]. Available: <http://arxiv.org/abs/1611.05708>
- [3] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," May 2018. *arXiv:1805.00655*. [Online]. Available: <http://arxiv.org/abs/1805.00655>
- [4] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Dec. 2015, pp. 4346–4354. doi: 10.1109/ICCV.2015.494.

- [5] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, “Learning human motion models for long-term predictions,” Apr. 2017, *arXiv:1704.02827*. [Online]. Available: <http://arxiv.org/abs/1704.02827>
- [6] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-RNN: Deep learning on spatio-temporal graphs,” Nov. 2015, *arXiv:1511.05298*. [Online]. Available: <http://arxiv.org/abs/1511.05298>
- [7] J. Martinez, M. J. Black, and J. Romero, “On human motion prediction using recurrent neural networks,” May 2017, *arXiv:1705.02445*. [Online]. Available: <http://arxiv.org/abs/1705.02445>
- [8] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, Aug. 2017, pp. 1243–1252. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3305381.3305510>
- [9] M. Dong and C. Xu, “On retrospecting human dynamics with attention,” Aug. 2019, pp. 708–714.
- [10] A. M. Lehmman, P. V. Gehler, and S. Nowozin, “Efficient nonlinear Markov models for human motion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1314–1321.
- [11] V. Pavlović, J. M. Rehg, and J. MacCormick, “Learning switching linear models of human motion,” in *Proc. 13th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2000, pp. 942–948. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3008751.3008888>
- [12] R. Urtasun, D. J. Fleet, A. Geiger, and J. Popović, T. J. Darrell, and N. D. Lawrence, “Topologically-constrained latent variable models,” in *Proc. 25th Int. Conf. Mach. Learn.* New York, NY, USA: ACM, 2008, pp. 1080–1087. doi: [10.1145/1390156.1390292](https://doi.org/10.1145/1390156.1390292).
- [13] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models for human motion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008. doi: [10.1109/TPAMI.2007.1167](https://doi.org/10.1109/TPAMI.2007.1167).
- [14] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh, “Bilinear spatiotemporal basis models,” *ACM Trans. Graph.*, vol. 31, no. 2, Apr. 2012, Art. no. 17. doi: [10.1145/2159516.2159523](https://doi.org/10.1145/2159516.2159523).
- [15] I. Sutskever, G. E. Hinton, and G. W. Taylor, “The recurrent temporal restricted boltzmann machine,” in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. New York, NY, USA: Curran Associates, 2009, pp. 1601–1608. [Online]. Available: <http://papers.nips.cc/paper/3567-the-recurrent-temporal-restricted-boltzmann-machine.pdf>
- [16] G. W. Taylor and G. E. Hinton, “Factored conditional restricted Boltzmann machines for modeling motion style,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.* New York, NY, USA: ACM, 2009, pp. 1025–1032. doi: [10.1145/1553374.1553505](https://doi.org/10.1145/1553374.1553505).
- [17] G. W. Taylor, G. E. Hinton, and S. T. Roweis, “Modeling human motion using binary latent variables,” in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 1345–1352. [Online]. Available: <http://papers.nips.cc/paper/3078-modeling-human-motion-using-binary-latent-variables.pdf>
- [18] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, “Dynamical binary latent variable models for 3D human pose tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 631–638.
- [19] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Sep. 2014, *arXiv:1409.0473*. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [20] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
- [21] A. Graves, “Generating sequences with recurrent neural networks,” Aug. 2013, *arXiv:1308.0850*. [Online]. Available: <http://arxiv.org/abs/1308.0850>
- [22] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [23] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” pp. 1171–1179, Dec. 2015. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2969239.2969370>
- [24] E. Barsoum, J. Kender, and Z. Liu, “HP-GAN: Probabilistic 3D human motion prediction via GAN,” *CoRR*, 2017. [Online]. Available: <http://arxiv.org/abs/1711.09561>
- [25] L.-Y. Gui, Y. Wang, X. Liang, and J. M. Moura, “Adversarial geometry-aware human motion prediction,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 786–803.
- [26] J. Butepage, M. Black, D. Kragic, and H. Kjellström, “Deep representation learning for human motion prediction and classification,” Jul. 2017, pp. 1591–1599.
- [27] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” Aug. 2015, *arXiv:1508.04025*. [Online]. Available: <http://arxiv.org/abs/1508.04025>
- [28] Y. Tang, L. Ma, W. Liu, and W.-S. Zheng, “Long-term human motion prediction by modeling motion context and enhancing motion dynamic,” in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 935–941. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3304415.3304548>
- [29] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, Jul. 2015, pp. 2048–2057. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3045118.3045336>
- [30] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2014, pp. 3104–3112. [Online]. Available: <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
- [31] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
- [32] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” Jun. 2014, *arXiv:1406.1078*. [Online]. Available: <http://arxiv.org/abs/1406.1078>



MINJING DONG received the bachelor's degree from the University of Sydney and the M.Phil. degree from the School of Computer Science, The University of Sydney. His research interests include human behavior analysis and network compression.



CHANG XU received the Ph.D. degree from Peking University, China. He is a Lecturer and an ARC DECRA Fellow of the School of Computer Science, The University of Sydney. His research interests include machine learning algorithms and related applications in computer vision. He has published over 50 papers in prestigious journals and top tier conferences, including the IEEE TPAMI, the IEEE TNNLS, the IEEE TIP, ICML, NIPS, IJCAI, and AAAI. He has received several paper awards including the Distinguished Paper Award from IJCAI 2018. He has been recognized as the Top Ten Distinguished Senior Program Chair of IJCAI 2017. He regularly served as (senior) program community member for many conferences.

...