Reciprocal collaboration network for 3D skeleton-based human motion prediction

Zhiquan He[a,b,c], Lujun Zhang[b], Wenming Cao[b*]

[a] College of Information Engineering, Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China; [b] Guangdong Multimedia Information Service Engineering Technology Research Center, Shenzhen University, Shenzhen, China; [c] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, China

ABSTRACT

Noticing that the human motion in both the positive and inverted time sequence are predictable, we design a reciprocal network which contains a forward and a backward prediction network to learn this information respectively, and then train the two networks in a collaborative and iterative way with the guidance of consistency constraint. Based on the reciprocal network, we establish an error compensation network to approximate the error of forward prediction network for making the forecast closer to the real pose. Extensive experiments show that our approach outperforms most recent methods in both short and long term motion predictions on Human 3.6M and CMU Mocap.

Keywords: Reciprocal network, error compensation, motion prediction

1. INTRODUCTION

With the given prior 3D pose sequence comprised of key joint and skeleton of body, 3D skeleton-based human motion prediction can forecast the future pose sequence, which plays a crucial role in numerous field[1-4]. For example, the 3D skeleton-based human motion prediction is vital for intelligent interaction like auto-driving[5] and pedestrian tracking[6-8].

For forecasting the future motion pose sequence, Fragkiadaki et al.[9] propose an Encoder-Recurrent-Decoder (ERD) network based on recurrent neural networks. Due to the relationships or constraints between different body parts of pose sequences, Li et al.[10] propose a multiscale graph neural networks to extract the characteristic relationships between different body parts. Guo[11] present a skeleton network (SkeNet) to learn local representations on different body components separately. Mao et al.[12] propose a simple feed-forward deep network for motion prediction, which takes both temporal smoothness and spatial dependencies among human body joints into account. Cui et al.[13] represent the skeletal pose as a novel dynamic graph and propose a deep generative model based on graph networks and adversarial learning.

Exiting research fully considers the temporal features of human motion pose sequence in positive temporal sequence, but does not consider the features of the inverted temporal sequence. According to Sun et al.[14], the inverted temporal pose sequence also contains the kinematics information of human body, which plays a key role in motion prediction. Let the time invert and image that person is moving backwards. The forward moving pose sequence follow the constraint of human kinematics. So do the backward moving pose sequence, since the only difference between them is that their directions of time, and they follow the consistency constraints.

Based on this, we propose a novel model, called Reciprocal Collaboration Network (RCN) for 3D skeleton-based human motion prediction. Our model is composed of three networks: a forward prediction network F which forecast the future motion pose in positive temporal sequence with the past motion pose, and a backward prediction network G which forecast the past motion pose in an inverse temporal sequence with the future motion, and an error compensation network to predict the error of F to make the result match the real pose better. Both the networks F and G are inverse operations to each other, which means that they satisfy a reciprocal constraint. After a reciprocal training stage, the networks F and G should have a strong consistency, which is the key to build the error compensation network (ECN). We experiment our method on CMU Mocap and Human 3.6M datasets, and the result demonstrates that our method is effective in human motion prediction.

[*] wmcao@szu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125063Q
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662586

2. RELATED WORK

In recent years, neural network had been increasingly recognized in human motion prediction. For example, Jain et al.[15] captures the rich interactions in the underlying spatial-temporal graph, and propose the Structural-RNN. Li et al.[10] presents a method to extract the structural features in different scale of human body component, and organizes a network forecasting the human pose based on GRU. In Li et al.[16], the problem of over-smoothing is solved by extracting comprehensive pose information from multiple spectral bands using graph scattering. Observing the repetition of human behaviour, Wei[17] utilize motion attention to capture the similarity from context.

Besides the approaches based on RNN, Li[18] present a method using the hierarchical structure of convolutional networks to capture spatial-temporal correlation, Aksan et al.[19] utilizes the transformer based architecture for generation of human motion modelling. Exiting work resolves the problem of motion prediction by capturing the pose feature in the positive time sequence. But the approaches above do not exploit the temporal features fully, Sun et al.[14] notices that the human motion pose, governed the constraints of human kinematics, are both forward predictable and backward predictable. In this work, we built a reciprocal network to extract the feature in both positive and inverted temporal sequence, and an error compensation based on the reciprocal network to correct the prediction of forward network further.

3. PROBLEM FORMULATION

Assuming that $X_{-T_p:0} = [X_{-T_p}, ..., X_0]$ represent the observed pose of length $T_p + 1$ in the past, where $X_i \in \mathbb{R}^{N \times D}$ with N joints and D = 3 dimension space data denote the motion pose at time i. The task of human motion prediction is to generate the future pose $X_{1:T_f} = [X_1, ..., X_{T_f}]$ of T_f length with the past pose observed $X_{-T_p:0}$.

In order to exploit the rich temporal features of pose, we add a backward prediction network to capture the information in an inverted temporal sequence, which can be another guidance of the forward prediction network.

4. METHOD OVERVIEW

Figure 1 shows the whole structure of our model. Backward network strength the forward network with a Consistency constraint, and ECN aim to improve forward network performance further by generating a tiny error compensation. To propose our reciprocal collaboration network (RCN), we introduce the network in four aspects in this section: reciprocal framework structure, the error compensation network, a new loss function of RCN and architecture of each network.

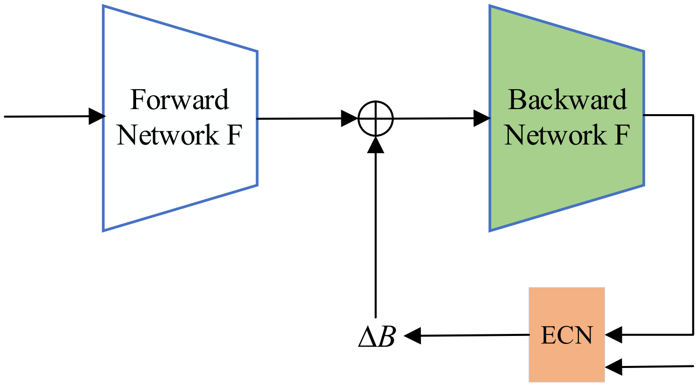

Figure 1. Whole structure of RCN.

4.1 Reciprocal framework structure

In order to capture the information of both the positive and 1nverted temporal sequence, our framework organizes a forward and a backward prediction network structure. As illustrated in Figure 2, both of the human motion in positive and inverted temporal sequence are predictable. The network F and G should be the same structure and they are pre-trained independently. It should be noted that the whole data, cover the future motion data, is usable during the training stage. So we can pre-train the network G by the same approach of network F.

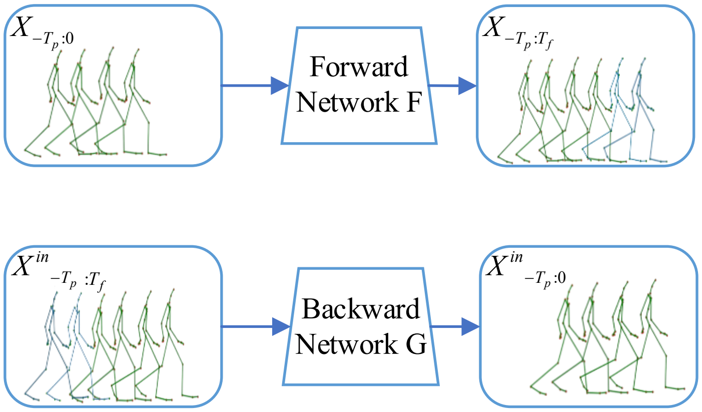

Figure 2. The pre-train stage of F and G. The pose described by green segment denotes observed pose, and the pose described by blue segment denotes predicted results. X represent the positive temporal and X^{in} represent the inverted one.

With the motion N observed past, network F predict the future motion $\widehat{M} = F(N)$, and network G predict the past motion $\widehat{N} = G(M)$ using the future motion M. If the network F and G learn well in pre-train stage, then they should satisfy the consistency constraints, which illustrated by

$$N \approx G(F(N)) \tag{1}$$

$$M \approx F(G(M)) \tag{2}$$

Utilizing the constrains above, both of the network F and G can be corrected further by each other. Specifically, when the network is trained, the network G can verify the accuracy of F. If the prediction $\widehat{M} = F(N)$ is incorrect, with this $\widehat{M}$ as input, G can't generate the inerrant motion that matches N due to the constrains (2). Likewise, when training the network G, F can also guide network G to correct. If G generate the incorrect forecast $\widehat{N}$, as the verified network, $\widehat{M}$ generated by F should have a large distance from M. These prediction distance of the verified network can also guide the trained network to correct their forecast.

Based on this, we can improve the performance of the incorrect network through a reciprocal collaboration learning. In reciprocal collaboration learning, we first train the network F and keep the network G which be used to verify F only. Then we train the network G and keep F. The forward network F and the backward network G are trained alternately and iteratively (Figure 3).

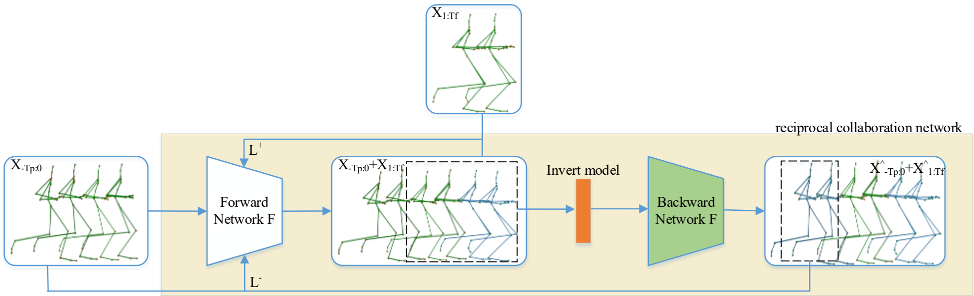

Figure 3. The forward network F learns in the reciprocal training stage.

4.2 Error compensation network

After the reciprocal train stage, both the network F and G have a strong consistency to each order. Base on this, we propose a novel approach to help the prediction of F match the ground truth better, which called the error compensation network (ECN).

We denote $\widehat{B}$ is the prediction of F, $\widehat{A}$ is the prediction of G, and $\Delta A = A - \widehat{A}$. As illustrated in Figure 4, using the error of G (ΔA)and the ground truth of F(B), ECN aim to predict ΔB , the error of network F, to help $\widehat{B}$ match B better.

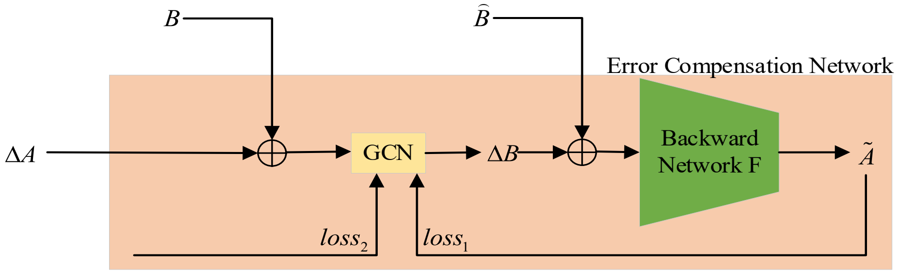

Figure 4. The architecture of RCN.

4.3 Loss function

Let X represent the truth pose in the positive sequence of time and X^{in} represent the inverted one. The forward network F aim to predict the future pose sequence $\widehat{X}_{1:T_f} = F(X_{-T_p:0})$. On the contrary, the backward network predict the past pose sequence $\widehat{X}^{in}_{1:T_f} = G(X^{in}{}_{-T_p:0})$, where the input $X^{in}{}_{-T_p:0}$ is a restructured sequence composes of $\widehat{X}_{1:T_f}$ and $X_{T_p+1-T_F:0}$ .In other word, using the predicted pose of F, we should restructure a $T_p + 1$ length sequence , as long as the input of F, to train G, which aims to keep the symmetry of F and G.

As is shown in Figure 2, the forward network F trained follow the guidance of L^+ and L^-, where the $L^+ = \widehat{X}_{1:T_f} - X_{1:T_f}$ and $L^- = \widehat{X}^{in}_{1:T_f} - X^{in}{}_{1:T_f}$. Combined the L^+ and L^- , the loss function of network F is:

$$L^{for} = \lambda L^+ + (1-\lambda)L^- \tag{3}$$

We design the loss function of network G in the same approach. The backward network G predict the past sequence $\widehat{X}^{in}_{1:T_f} = G(X^{in}_{-T_p})$, and F predict the future sequence $X_{1:T_f} = F(X_{-T_p:0})$ with the restructured sequence. The backward network G trained follow the guidance of L^- and L^+ , where $L^- = \widehat{X}^{in}_{1:T_f} - X^{in}{}_{1:T_f}$ and $L^+ = \widehat{X}_{1:T_f} - X_{1:T_f}$, the loss function of network G is:

$$L^{back} = \lambda L^- + (1-\lambda)L^+ \tag{4}$$

In reciprocal training stage, we train the network F and G alternately with L^{for} and L^{back} , which improves the performance both of two network.

ECN Loss Function: We design double loss to train the error compensation network. As shown in Figure 4, $\tilde{A}$ denotes the prediction of G corrected by ΔB . Hoping the network F have a better forecast, $\tilde{A}$ should match A better than $\widehat{A}$. Then we design $loss_1 = A - \tilde{A}$ and $loss_2 = \left|A - \tilde{A}\right| - \left|A - \widehat{A}\right|$. The loss function is:

$$Loss = \lambda_1 loss_1 + \lambda_2 loss_2 \tag{5}$$

4.4 Network Architectures

Our forward and backward prediction network are the same structure and adapt the architecture from Maosen Li et al.[10] who have shown impressive performance for features extraction from multiscale human body component. Both of two network follows the encoder-decoder model, where encoder contain the cascaded multiscale graph computational blocks

(MGCU) to extract and fuse the feature in different scale of body component, and decoder is organized base on GRU. And ECN is based on GCN.

5. EXPERIMENTS

5.1 Datasets

CMU Mocap datasets has 5 main categories data of motion: ‘human interaction’, ‘interaction with environment’, ‘locomotion’, ‘physical activities & sports’, and ‘situations & scenarios’. Be consistent with Li et al.[18], we select 8 action: ‘basketball’, ‘basketball signal’, ‘directing traffic’, ‘jumping’, ‘running’, ‘soccer’, ‘walking’ and ‘washing window’, where each motion data consists of 38 joint and we preserve 26 useful joint.

Human 3.6m datasets has 15 different classes of motion performed by 7 actors, where each subject contains 32 joint and discard 10 redundant joints. Be consistent with Li et al.[10], we train the model on 6 subjects and test on 5th subject.

5.2 Evaluation Metrics

We consider a mean angle error (MAE) to evaluate our model. Let α , β , γ denote the euler angles between the adjacent bones, the MAE is evaluated as:

$$MAE = \frac{1}{m-1}\sum\left\|(\alpha-\widehat{\alpha})+(\beta-\widehat{\beta})+(\gamma-\widehat{\gamma})\right\|_2 \tag{6}$$

5.3 Comparison to state-of-the-art methods

We compare our model to last method[10, 12, 18-20] on the two datasets. Short-term prediction predicts the future pose within 500 milliseconds. As shown in the Tables 1 and 2, we compare our approach with recent work at “80ms”, “160ms”, “320ms” and “400ms” and consistent to Li et al.[10]. In Human3.6M datasets, our approach has advantage in action “Posing”, “purchases”, “Sitting down”, “Waiting” and “Walking Together” and obtains competitive results on others. In CMU Mocap datasets, our method outperforms recent approach in “Basketball”, “Basketball Signal”, “running” and “working”

Table 1. MAE of different methods for short-term prediction on Human3.6M. We also present the RCN without ECN model.

Motion	**Walking**				**Eating**				**Smoking**				**Discussion**			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU[12]	0.29	0.49	0.71	0.78	0.25	0.42	0.68	0.83	0.32	0.6	1	1.11	0.31	0.69	1.03	1.12
CEM[18]	0.33	0.54	0.68	0.73	0.22	0.36	0.58	0.71	0.26	0.49	0.96	0.92	0.32	0.67	0.94	1.01
Traj-GCN[12]	0.18	0.31	0.49	0.56	0.16	0.29	0.5	0.62	0.22	0.41	0.86	0.8	0.2	0.51	0.77	0.85
DMGNN[10]	0.18	0.31	0.49	0.58	0.17	0.3	0.49	0.59	0.22	0.39	0.81	0.77	0.26	0.65	0.92	0.99
RCN(-ECN)	0.19	0.31	0.54	0.57	0.16	0.28	0.47	0.59	0.23	0.41	0.86	0.78	0.30	0.68	0.95	1.00
RCN (Ours)	0.19	0.31	0.52	0.56	0.16	0.28	0.45	0.58	0.22	0.41	0.84	0.77	0.29	0.67	0.97	1.03
Motion	Directions				Greeting				Phoning				Posing			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU[20]	0.41	0.64	0.8	0.92	0.57	0.83	1.45	1.6	0.59	1.06	1.45	1.6	0.45	0.85	1.34	1.56
CEM[18]	0.39	0.6	0.8	0.91	0.51	0.82	1.21	1.38	0.59	1.13	1.51	1.65	0.29	0.6	1.12	1.37
Traj-GCN[12]	0.26	0.45	0.7	0.79	0.35	0.61	0.96	1.13	0.53	1.02	1.32	1.45	0.23	0.54	1.26	1.38
DMGNN[10]	0.25	0.44	0.65	0.71	0.36	0.61	0.94	1.12	0.52	0.97	1.29	1.43	0.2	0.46	1.06	1.34
RCN(-ECN)	0.25	0.45	0.67	0.73	0.38	0.60	0.96	1.12	0.52	1.00	1.30	1.44	0.22	0.45	0.99	1.2

RCN (Ours)	0.25	0.44	0.67	0.73	0.38	0.60	0.94	1.11	0.52	1.00	1.30	1.43	0.2	0.43	0.97	1.18
Motion	Purchases				Sitting				Sitting Down				Taking Photo			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU[20]	0.58	0.79	1.08	1.15	0.41	0.68	1.12	1.33	0.47	0.88	1.37	1.54	0.28	0.57	0.9	1.02
CEM[18]	0.63	0.91	1.19	1.29	0.39	0.61	1.02	1.18	0.41	0.78	1.16	1.31	0.23	0.49	0.88	1.06
Traj-GCN[12]	0.42	0.66	1.04	1.12	0.29	0.45	0.82	0.97	0.3	0.63	0.89	1.01	0.15	0.36	0.59	0.72
DMGNN[10]	0.41	0.61	1.05	1.14	0.26	0.42	0.76	0.97	0.32	0.65	0.93	1.05	0.15	0.34	0.58	0.71
RCN(-ECN)	0.46	0.68	1.03	1.11	0.25	0.42	0.79	0.97	0.31	0.63	0.93	1.01	0.16	0.38	0.62	0.77
RCN (Ours)	0.46	0.65	1.01	1.09	0.25	0.42	0.81	0.99	0.3	0.62	0.90	1.01	0.15	0.36	0.62	0.75
Motion	Waiting				Walking Dog				Walking Together							
millisecond	80	160	320	400	80	160	320	400	80	160	320	400				
ResGRU[20]	0.32	0.63	1.07	1.26	0.52	0.89	1.25	1.4	0.27	0.53	0.74	0.79				
CEM[18]	0.3	0.62	1.09	1.3	0.59	1	1.32	1.44	0.27	0.52	0.71	0.74				
Traj-GCN[12]	0.23	0.50	0.92	1.15	0.46	0.8	1.12	1.3	0.15	0.35	0.52	0.57				
DMGNN[10]	0.22	0.49	0.88	1.1	0.42	0.72	1.16	1.34	0.15	0.33	0.50	0.57				
RCN(-ECN)	0.23	0.50	0.94	1.11	0.41	0.77	1.15	1.35	0.15	0.33	0.50	0.55				
RCN (Ours)	0.22	0.49	0.94	1.07	0.43	0.78	1.17	1.33	0.15	0.33	0.49	0.54				

Table 2. MAE of different method in 8 actions of CMU Mocap. We present the MAE at both short and long term.

Motion	**Basketball**					**Basketball Signal**					**Directing Traffic**					**Jumping**				
millisecond	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ResGRU[20]	0.49	0.77	1.26	1.45	1.77	0.42	0.76	1.33	1.54	2.17	0.31	0.58	0.94	1.1	2.06	0.57	0.86	1.76	2.03	2.42
CEM[18]	0.36	0.62	1.07	1.17	1.95	0.33	0.62	1.05	1.23	1.98	0.26	0.58	0.91	1.04	2.08	0.38	0.6	1.36	1.58	2.05
Traj-GCN[12]	0.33	0.52	0.89	1.06	1.71	0.11	0.2	0.41	0.53	1	0.15	0.32	0.52	0.6	2	0.31	0.49	1.23	1.39	1.8
DMGNN[10]	0.3	0.46	0.89	1.11	1.66	0.1	0.17	0.31	0.41	1.26	0.15	0.3	0.57	0.72	1.98	0.37	0.65	1.49	1.71	1.79
RCN(-ECN)	0.3	0.47	0.88	1.10	1.66	0.1	0.17	0.31	0.41	0.89	0.15	0.32	0.57	0.72	1.98	0.35	0.62	1.48	1.69	1.77
RCN (Ours)	0.29	0.47	0.86	1.05	1.69	0.09	0.17	0.32	0.41	0.81	0.16	0.37	0.59	0.71	2	0.35	0.61	1.45	1.68	1.72
Motion	Running					Soccer					Walking					Washing Window				
millisecond	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
ResGRU[20]	0.32	0.48	0.65	0.74	1	0.29	0.5	0.87	0.98	1.73	0.35	0.45	0.59	0.64	0.88	0.31	0.47	0.74	0.93	1.37
CEM[18]	0.28	0.43	0.54	0.57	0.69	0.28	0.48	0.79	0.9	1.58	0.35	0.44	0.46	0.51	0.77	0.3	0.47	0.79	1	1.39
Traj-GCN[12]	0.33	0.55	0.73	0.74	0.95	0.18	0.29	0.61	0.71	1.4	0.33	0.45	0.49	0.53	0.61	0.22	0.33	0.57	0.75	1.2
DMGNN[10]	0.19	0.31	0.47	0.49	0.64	0.22	0.32	0.79	0.91	1.54	0.3	0.34	0.38	0.43	0.6	0.2	0.27	0.62	0.81	1.09
RCN(-ECN)	0.19	0.29	0.45	0.9	0.64	0.20	0.33	0.65	0.77	1.51	0.3	0.33	0.38	0.45	0.61	0.2	0.28	0.59	0.80	1.19
RCN (Ours)	0.19	0.29	0.44	0.47	0.66	0.20	0.37	0.65	0.73	1.49	0.3	0.33	0.36	0.39	0.56	0.2	0.27	0.6	0.78	1.15

Long-term prediction aims to forecast the pose over 500 milliseconds. We compare our result at “1000ms” in CMU Mocap, and the result at “560ms” and “1000ms” in Human3.6M. Tables 2 and 3 shows the MAE of various approach. In CMU Mocap, our approach has an accurate prediction on action “Basketball Signal” and have a competitive result on action “Jumping” and “Walking”. In human3.6M, our model has a lowest MAE in action “Discussion”.

Table 3. MAE of various method on the 4 representative actions of H3.6M dataset for long-term prediction.

Motion	**Walking**		**Eating**		**Smoking**		**Discussion**	
Millisecond	560	1000	560	1000	560	1000	560	1000
ResGRU[20]	0.93	1.03	0.95	1.08	1.25	1.5	1.43	1.69
CEM[18]	0.98	0.92	1.01	1.24	0.97	1.62	1.56	1.86
Traj-GCN[12]	0.65	0.67	0.76	1.12	0.87	1.57	1.33	1.7
DMGNN[10]	0.66	0.75	0.74	1.14	0.83	1.52	1.33	1.45
RCN(-ECN)	0.69	0.92	0.79	1.22	0.85	1.56	1.31	1.41
RCN	0.69	0.89	0.82	1.24	0.83	1.55	1.31	1.38

6. CONCLUSIONS

Considering the human motion in an inverted time sequence also contains kinesiology information, we propose a reciprocal network which contains a forward network and a backward network for 3D skeleton-based human motion prediction, and train each network alternately and iteratively based on consistency constraints. This leads to our network being equipped with a model with rich kinesiology information in different directions of time sequence. Noticing that the consistency of the forward and backward prediction network, we design an error compensation network to generate a prediction error as a compensation of the forward network, which improves the performance of reciprocal network effectively. The result shows that our approach is effective both in short and long term motion prediction on public benchmark datasets.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under grants 61971290, 61771322, 61871186, and the Fundamental Research Foundation of Shenzhen under Grant JCYJ20190808160815125.

REFERENCES

[1] Koppula, H., and Saxena, A., “Learning spatiotemporal structure from RGB-D videos for human activity detection and anticipation,” PMLR, 28(3), 792-800(2013).

[2] Koppula, H., and Saxena, A., “Anticipating human activities using object affordances for reactive robotic response,” IEEE. Trans., 38(1), 14-29(2016).

[3] Gui, L., Zhang, K., Wang, Y., Liang, X., Moura, J. and Veloso, M. “Teaching robots to predict human motion,” IROS, 562-567(2018).

[4] Huang, D. and Kitani, K., “Action-reaction: Forecasting the dynamics of human interaction,” European Conference on ECCV, Springer, 489-504(2014).

[5] Chen, S., Liu, B., Feng, C., VallespiGonzalez, C. and Wellington, C., “3D point cloud processing and learning for autonomous driving,” IEEE. ISO4, 38(1), 68-86(2020).

[6] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Li, F. and Savarese, S., “Social LSTM: Human trajectory prediction in crowded spaces,” CVPR, 961-971(2016).

[7] Gupta, A., Martinez, J., Little, J., and Woodham, R., “3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding,” CVPR, 2061-2068(2014).

[8] Bhattacharyya, A., Fritz, M. and Schiele, B., “Long-term on-board prediction of people in traffic scenes under uncertainty,” CVPR, 4194-4202(2018).

[9] Fragkiadaki, K., Levine, S., Felsen, P. and Malik, J., “Recurrent network models for human dynamics,” ICCV, 4346-4354(2015).

[10] Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y. and Tian, Q., “Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction,” CVPR, 214-22(2020).

[11] Guo, X. and Choi, J., “Human motion prediction via learning local structure representations and temporal dependencies,” AAAI, 2580-2587(2019).

[12] Mao, W., Liu, M., Salzmann, M. and Li, H., “Learning trajectory dependencies for human motion prediction,” ICCV, 9489-9497(2019).

[13] Cui, Q., Sun, H. and Yang, F., “Learning dynamic relationships for 3d human motion prediction,” CVPR, 6518-6526(2020).

[14] Sun, H., Zhao, Z., Yin, Z. and He, Z., “Reciprocal twin networks for pedestrian motion learning and future path prediction,” IEEE. Trans., 32(3), 1483-1497(2021).

[15] Jain, A., Zamir, A. R., Savarese, S. and Saxena, A., “Structural-RNN: Deep learning on spatio-temporal graphs,” CVPR, 5308-5317(2016).

[16] Li, M., Chen, S., Liu, Z., Zhang, Z., Xie, L., Tian, Q., and Zhang, Y., “Skeleton graph scattering networks for 3d skeleton-based human motion prediction,” ICCV, 854-864(2021).

[17] Mao, W., Liu, M. and Salzemann, M., “History repeats itself: Human motion prediction via motion attention,” ICCV, 474-489(2020).

[18] Li, C., Zhang, Z., Lee W. S., and Lee, G. H., “Convolutional sequence to sequence model for human dynamic,” CVPR, 5226-5234(2018).

[19] Aksan, E., Cao, P., Kaufmann, M. and Hilliges, O., “A spatiotemporal transformer for 3D human motion prediction,” 3DV, 565-574(2021).

[20] Martinez, J., Black, M. and Romero, J., “On human motion prediction using recurrent neural networks,” CVPR, 4674-4683(2017).