Stacked residual blocks based encoder–decoder framework for human motion prediction

Xiaoli Liu¹, Jianqin Yin¹

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing, People’s Republic of China

Abstract: Human motion prediction is an important and challenging task in computer vision with various applications. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been proposed to address this challenging task. However, RNNs exhibit their limitations on long-term temporal modelling and spatial modelling of motion signals. CNNs show their inflexible spatial and temporal modelling capability that mainly depends on a large convolutional kernel and the stride of convolutional operation. Moreover, those methods predict multiple future poses recursively, which easily suffer from noise accumulation. The authors present a new encoder–decoder framework based on the residual convolutional block with a small filter to predict future human poses, which can flexibly capture the hierarchical spatial and temporal representation of the human motion signals from the motion capture sensor. Specifically, the encoder is stacked by multiple residual convolutional blocks to hierarchically encode the spatio-temporal features of previous poses. To automatically reconstruct the spatial and temporal information of future poses in a non-recursive manner, which can avoid noise accumulation that differs from prior works. Experimental results show that the proposed method outperforms baselines on the Human3.6M dataset, which shows the effectiveness of the proposed method. The code is available at https://github.com/lily2lab/residual_prediction_network.

1 Introduction

Human motion prediction is attracting increasing attention in computer vision and robotics with its wide applications, such as human–machine cooperation [1, 2], intelligent security [3], and so on [4, 5]. Predicting future dynamics of the human body plays an important role to ensure the safety of human life and property. For example, if robots can predict that people are likely to fall in the near future, people can avoid danger with a great probability. As is shown in Fig. 1, the human motion sequence can be considered as the signals of a group points evolving with time. Therefore, human motion prediction is a signal prediction problem that predicts the future signals of these skeletal points based on the currently received signals.

Previous studies utilise a recurrent neural network (RNN) to construct their encoder–decoder frameworks to predict future human poses [3, 6–8]. For example, Martinez et al. [8] and Gui et al. [3] proposed their encoder–decoder frameworks entirely based on the gated recurrent unit (GRU) [9]. As is known to all, RNN is shown in Fig. 1, the human motion sequence can be considered as a sequence of the current step as the input of the next step, the error of each predictive pose will be accumulated across time step by step. Therefore, the predictive performance of the final network [15]. Moreover, the work in [14] has shown that a simple convolutional architecture outperforms canonical recurrent networks such as LSTMs. Therefore, in this paper, we propose to construct our encoder–decoder framework using the CNN.

Most existing methods for human motion prediction easily suffered from noise accumulation [3, 8, 10]. Since these models predicted multiple poses recursively by considering the output of the current step as the input of the next step, the error of each predictive pose will be accumulated across time step by step. Therefore, the predictive performance of the later timestamps will worse greatly.

Different from previous works, we present a new encoder–decoder framework, stacked residual blocks based encoder–decoder, to predict all future poses within a short interval in a non-recursive way. Specifically, an encoder is built based on residual convolutional block, which can capture the strong correlations of joints of one limb by augmenting local features from lower layers through residual connection. In this encoder, multiple residual convolutional blocks are stacked to enlarge both the spatial and temporal receptive field layer by layer, which can better model the hierarchical spatio-temporal representation of previous poses. To reconstruct the spatial and temporal information of future poses, a decoder is constructed with two fully connected (FC) layers, which can automatically capture the global spatial and temporal features and flexibly predict multiple future poses at one step.

Our contributions are summarised as follows:

- **Abstract:** Human motion prediction is an important and challenging task in computer vision with various applications. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been proposed to address this challenging task. However, RNNs exhibit their limitations on long-term temporal modelling and spatial modelling of motion signals. CNNs show their inflexible spatial and temporal modelling capability that mainly depends on a large convolutional kernel and the stride of convolutional operation. Moreover, those methods predict multiple future poses recursively, which easily suffer from noise accumulation. The authors present a new encoder–decoder framework based on the residual convolutional block with a small filter to predict future human poses, which can flexibly capture the hierarchical spatial and temporal representation of the human motion signals from the motion capture sensor. Specifically, the encoder is stacked by multiple residual convolutional blocks to hierarchically encode the spatio-temporal features of previous poses. To automatically reconstruct the spatial and temporal information of future poses in a non-recursive manner, which can avoid noise accumulation that differs from prior works. Experimental results show that the proposed method outperforms baselines on the Human3.6M dataset, which shows the effectiveness of the proposed method. The code is available at https://github.com/lily2lab/residual_prediction_network.

- **Introduction:** Human motion prediction is attracting increasing attention in computer vision and robotics with its wide applications, such as human–machine cooperation [1, 2], intelligent security [3], and so on [4, 5]. Predicting future dynamics of the human body plays an important role to ensure the safety of human life and property. Therefore, these models cannot capture spatial features efficiently. Convolutional neural network (CNN)-based encoder–decoder framework. To address this limitation of RNNs, some researchers proposed their models based on the CNN [10–14]. For example, Li et al. [10] proposed a new sequence-to-sequence model based on CNN, which makes their model capture both spatial and temporal features effectively. However, their spatial and long-term temporal dependencies modelling abilities heavily rely on a large convolutional kernel and the stride of convolutional operation. However, it has been shown that a smaller convolutional filter is better and more flexible for the performance of the final network [15]. Moreover, the work in [14] has shown that a simple convolutional architecture outperforms canonical recurrent networks such as LSTMs. Therefore, in this paper, we propose to construct our encoder–decoder framework using the CNN.

- **Stacked residual blocks:** In the encoder, multiple residual convolutional blocks are stacked to enlarge both the spatial and temporal receptive field layer by layer, which can better model the hierarchical spatio-temporal representation of previous poses. To reconstruct the spatial and temporal information of future poses, a decoder is constructed with two fully connected (FC) layers, which can automatically capture the global spatial and temporal features and flexibly predict multiple future poses at one step.
A new stacked residual convolutional blocks based encoder–decoder is proposed to predict the future motion sequence of the human body, which can capture both spatial and temporal features efficiently. Moreover, the proposed framework predicts multiple future poses in a non-recursive manner, which can avoid error accumulation to a great extent.

Experimental results show that our proposed method achieves superior performance compared with baselines on benchmark dataset (Human3.6M), which demonstrates the effectiveness of our proposed method.

2 Related work

Human motion prediction is one of the sequence-to-sequence modelling problems, and many kinds of literature were presented to study this problem, which can be summarised as follows:

Most sequence-to-sequence models have been proposed based on RNN to predict future human dynamics [3, 6–8, 16–21], such as GRU [9] and long short-term memory (LSTM) [22]. Early works modelled the spatial and temporal information of the previous human motion sequences entirely using the recurrent units in GRU or LSTM [3, 6, 8]. Fragkiadaki et al. [6] used LSTMs to construct their prediction model, encoder–recurrent–decoder, which jointly learn representations and dynamics of the human motion. Martinez et al. [8] proposed a GRU-based residual architecture to model the velocities information of the future human motion sequence, which addresses the problem of discontinuity in the first prediction.

Although traditional RNNs have shown their power in short-term modelling, they are internally difficult to model the spatial information of the human body efficiently and capture the long-term dynamics of the human motion using their recurrent units. Therefore, some works presented to model long-term information by using RNNs hierarchically [17] or using multi-level frameworks [16]. Specifically, some works presented to model the long-term information by using RNNs hierarchically [17] or using multi-level frameworks [16]. Chiu et al. [17] presented a hierarchical model, triangular-prism RNN, to capture the latent hierarchical temporal structure of the human motion sequence by using LSTMs with different time-scales. Gopalakrishnan et al. [16] proposed a two-level processing architecture to predict future motion sequences. The lower-level RNN served as a learnable noise process to learn a sequence of ‘guide vectors’ to enhance the input features, and the bottom-level RNN was used to predict the future poses with the enhanced input features. Moreover, other works addressed the limitation of RNNs in spatial modelling to some extent [7, 20, 21]. Jain et al. [7] proposed a structural-RNN architecture through high-level spatio-temporal graphs and RNNs, which can capture rich interactions between different limbs of the human body. Liu et al. [21] captured the anatomical constraints of the human body using Lie algebra representation and captured the temporal context with a novel hierarchical motion recurrent network by using LSTMs hierarchically.

Another type of sequence-to-sequence model is proposed based on the CNN [10–14]. Pavllo et al. [12] proposed QuaterNet to represent rotations with quaternions and model the temporal information using dilated convolutions, which can model both short-term and long-term information efficiently. Li et al. [10] proposed a convolutional sequence-to-sequence model to predict future human dynamics. The authors modelled the local features of the human body by designing a rectangle convolutional kernel and modelled the temporal information for short-term or long-term modelling by mapping a window of frames into a hidden variable using a convolutional encoder module. However, the window size for short-term modelling is designed by manual, which is inflexible.

Some works are proposed for human motion prediction by combining generative adversarial network (GAN) with RNN or CNN [23–26]. Barsoum et al. [26] proposed a novel sequence-to-sequence probabilistic prediction model to predict various future motion sequences given a certain input motion sequence with a different vector sampling from a random distribution. Kundu et al. [24] proposed a probabilistic generative model, bidirectional human motion prediction GAN (BiHMP-GAN), to address the prediction of uncertain future motion dynamics by taking into account the inherent stochasticity. Different from human prediction (HP)-GAN [26], BiHMP-GAN [24] incorporated a novel conditional discriminator to discriminate the true poses from the predicted ones and regress the random extrinsic vector, which enables the discriminator to avoid the problem of mode collapse and enforce direct content loss on the output of the decoder. So that BiHMP-GAN can achieve superior performance. Hernandez et al. [25] proposed a GAN to forecast the future motion sequence. The proposed method predicted more realistic results by modelling the essence and semantics of the human motion guided by three independent discriminators, which encourage the generation of motion sequences with a similar frequency distribution.

To sum up, although great success has been made, most of these models predicted multiple frames in a recursive manner, which easily suffer from error accumulations.

3 Methodology

3.1 Problem formulation

Standard methods of human motion prediction are proposed based on the encoder–decoder framework. Given input pose signals \( S = \{x_1, x_2, \ldots, x_T\} \) with length \( T \), and its corresponding output pose signals \( \tilde{S} = \{x_{T+1}, x_{T+2}, \ldots, x_{T+N}\} \) with length \( N \). Here, \( x_i \) is the \( i \)th pose of the human motion sequence. The encoder–decoder framework can be considered as two learning processes: the encoder first encodes previous poses into a latent variable \( z \), which can be formulated as (1). Then, the decoder maps future poses from the latent variable \( z \), which can be formulated as (2).

\[
\begin{align*}
    z &= f(S) = f(x_1, x_2, \ldots, x_T) \\
    \tilde{S} &= g(z)
\end{align*}
\]

where \( f(\cdot) \) and \( g(\cdot) \) denote two mappings that can be learned by encoder and decoder, respectively.

3.2 Stacked residual blocks-based encoder–decoder

In this paper, our proposed method follows the encoder–decoder framework and the overall architecture is shown in Fig. 2. Firstly, ‘data processing’ aims to convert input pose signals into a 3D tensor denoted at the top of Fig. 2. Then, ‘encoder’ is built with six residual convolutional blocks to encode the latent spatio-temporal representation of the input signals. Next, the output tensor of the encoder is flattened into a vector that can be processed by the FC.
layer. Finally, ‘decoder’ is built with two FC layers to reconstruct the spatial and temporal information of the future pose signals from the latent representation of the input pose signals. We will discuss ‘data processing’, ‘encoder’ and ‘decoder’, respectively, as follows:

(i) Data processing: due to the physical constraints of the human body in the process of the human movement, joints of the same limb have a strong correlation. To conveniently model those correlations using convolutional operation, during data processing, joints of the same limb are organised in the adjacent areas. In experiments, the final order of limbs is: ‘left arm’, ‘right arm’, ‘trunk’, ‘left leg’, and ‘right leg’.

(ii) Encoder: the spatial and temporal features of the human motion sequence are captured simultaneously by covering both space and time dimensions using a convolutional kernel. With this special organised manner of the input tensor, spatio-temporal features of some joints (i.e. local features) are captured at lower layers, and spatio-temporal features of more joints are captured at deeper layers. To model the strong correlations of joints of the same limb, local features should be enhanced. Therefore, as is shown in Fig. 2, the residual convolutional block [27] is introduced. With the residual connection between the lower layer and the deeper layer, ‘merging shallow features with deep features’, so that: (a) the local features can be further enhanced to capture the strong correlations of joints of one limb; (b) the receptive field can be enlarged by layer to model the global features of the human body and capture the hierarchical representation of the input signals. With those elegant designs, the encoder can better capture the spatial and temporal features of the input pose signals.

(iii) Decoder: to predict future human motion, the decoder needs to reconstruct the spatial and temporal information of future poses from the latent representation of previous poses. To avoid noise accumulation, all future poses needed to predict should be forecasted in a non-recursive manner. Therefore, the FC layer is a suitable solution, which can automatically model the global spatio-temporal features of human poses and flexibly address the prediction of multiple future poses in a non-recursive manner. As is shown in Fig. 2, our decoder is built with two FC layers, which reconstructs the global spatial and temporal information of all future poses within a short interval at one step.

4 Experiments

| Table 1 | Quantitative results on H3.6M. The results of ‘Res[8]’ and ‘Conv[11]’ are given in [30]. ‘ms’ denotes milliseconds |
|----------------|-----------------|-----------------|-----------------|
|               | Walking         | Eating          | Smoking         |
| ms            | 80              | 160             | 320             | 400             | 80              | 160             | 320             | 400             |
| Res[8]        | 24              | 40              | 63              | 71              | 18              | 35              | 72              | 88              | 20              | 37              | 62              | 74              |
| Conv[11]      | 17              | 31              | 54              | 62              | 14              | 26              | 53              | 63              | 11              | 21              | 33              | 38              |
| ours          | 12              | 19              | 36              | 40              | 14              | 24              | 45              | 53              | 22              | 29              | 45              | 53              |

| ms            | Discussion      | Directions      | Greeting       |
|----------------|-----------------|-----------------|----------------|
| 80              | 160             | 320             | 400             | 80              | 160             | 320             | 400             |
| Res[8]        | 32              | 61              | 96              | 104             | 37              | 56              | 82              | 97              | 38              | 74              | 139             | 159             |
| Conv[11]      | 19              | 39              | 68              | 76              | 22              | 37              | 60              | 73              | 25              | 46              | 90              | 103             |
| ours          | 16              | 29              | 48              | 53              | 16              | 26              | 56              | 68              | 17              | 32              | 73              | 89              |

| ms            | Phoning         | Posing          | Purchases      |
|----------------|-----------------|-----------------|----------------|
| 80              | 160             | 320             | 400             | 80              | 160             | 320             | 400             | 80              | 160             | 320             | 400             |
| Res[8]        | 26              | 44              | 74              | 84              | 28              | 55              | 131             | 161             | 41              | 72              | 104             | 110             |
| Conv[11]      | 17              | 30              | 53              | 61              | 16              | 36              | 86              | 106             | 29              | 55              | 82              | 93              |
| ours          | 23              | 31              | 52              | 60              | 26              | 40              | 87              | 107             | 28              | 48              | 80              | 89              |

4.1 Dataset and implementation details

Dataset: Human3.6M (H3.6M) [28, 29] is the current largest dataset for human motion prediction, which consists of 15 actions performed by seven professional actors, such as walking, eating, and smoking. The data captures from 15 sensors, including 4 digital video cameras, 1 time-of-flight sensor, and 10 motion capture cameras. The dataset provides both joint positions and joint angle skeletal representations. The joint positions are represented in a 3D coordinate system, which is obtained from the joint angles data provided by Vicon's skeleton fitting procedure by applying forward kinematics on the skeleton of the human body. The human pose is represented by 32 joints.

Implementation details: following the same experimental settings as baselines [8, 10], we validate our model on the sequences of subject 11, test on the sequences of subject 5, and train on the sequences of the remaining 5 subjects. In our proposed network, the channels of all convolutional layers are set to 64. Also, the output dimension of each layer in the encoder is set to 300 and (N_f × N_j × N_k), respectively, where N_f equals to the number of predictive poses, N_j is the number of joints of the human body, and N_k is the dimension of joints. In experiments, we use the previous ten frames to predict the future ten frames. All experiments are conducted using 3D position data. All models are implemented by TensorFlow and trained with Adam optimiser, and the initial learning rate is 0.0001. The mean per joint position error (MPJPE) proposed in [28] is used as our loss. Consistent with baselines, MPJPE is selected as our metric, which can be denoted as

\[ E = \frac{1}{N \times O} \sum_{j=1}^{O} \sum_{f=1}^{N} \| \hat{x}_{f,j} - x_{f,j} \| \]

where \( x_{f,j} \) is the ground-truth one and \( \hat{x}_{f,j} \) is the predictive one.

4.2 Quantitative results

To evaluate the effectiveness of our proposed method, experiments were carried out on H3.6M. Table 1 reports the detailed results of 15 actions and their average performance over all actions. For a more fair comparison, the results of Res[8] and Conv[11] used in Table 1 are reported in [30]. As is shown in Table 1, the errors of our method are decreased significantly in most cases such as ‘walking’, ‘eating’, ‘discussion’, ‘waiting’, and ‘walking together’, which demonstrates the effectiveness of our proposed method.
Taking action ‘walking’ as an example, (i) compared with the RNN-based baseline ‘Res[8]’, the errors of ‘80’, ‘160’, ‘320’, and ‘400 ms’ decrease by 12, 21, 27, and 31, respectively. Since our proposed CNN model is built with temporal convolution, incorporating our special organised input tensor of the skeletal convolutional kernel designed by manual, which cannot model the spatial structure information of the human body efficiently. ‘Res[8]’ built its model entirely based on GRU, which cannot capture the spatial structure of the human body efficiently. Therefore, compared with the RNN’s baseline [8], our model can significantly improve the predictive performance. (ii) Compared with the CNN-based baseline ‘Conv[11]’, the errors of ‘80’, ‘160’, ‘320’, and ‘400 ms’ decrease by 5, 12, 18, and 22, respectively. Our residual convolutional encoder–decoder model enables our network to capture the strong correlations of the joints of one limb, which can better describe the spatial and temporal information efficiently. ‘Res[8]’ modelled the spatial structure information of the human body and long-term temporal information of the human motions can be captured by stacking multiple residual convolutional blocks to enlarge the receptive field from both spatial and temporal dimensions. However, ‘Conv[11]’ modelled the spatial and temporal information from the previous human motion sequence using a plain CNN, and captured the spatial correlations of joints from different limbs with a large convolutional kernel designed by manual, which cannot model the spatial structure information of the human body efficiently. So that, compared with the CNN’s baseline [10], our residual CNN’s framework can better model the spatio-temporal information of the human motion sequence. Moreover, our network predicts a window of future poses in a non-recursive manner, while ‘Conv[11]’ predicted multiple future poses recursively. So that the later predictive poses of our model can avoid the inference of the noise from the early predictive poses efficiently. This may be one of the reasons for our better performance at later timestamps such as ‘320’ and ‘400 ms’ than other baselines.

Moreover, there shows a similar performance on average. Also, compared with baselines, our proposed method achieves the best performance on average, especially for long-term prediction, which evidence the effectiveness of our proposed method again. Specifically, compared with ‘Res[8]’, the errors of ‘80’ and ‘160 ms’ decrease by 10 and 24, respectively, while the errors of ‘320’ and ‘400 ms’ decrease by 39 and 44, respectively; compared with ‘Conv[11]’, the errors of ‘160’, ‘320’, and ‘400 ms’ decrease by 5, 7, and 8, respectively, and the error of ‘80’ is slightly worse, but the error is very close.

In summary, compared with both RNN’s [8] baseline and CNN’s [10] baseline, our proposed method achieves superior predictive performance, which may benefit from two folds: (a) our residual convolutional encoder enables the network to capture the hierarchical representation of the input pose signals and the strong correlations of joints of one limb, which can better describe the spatial and temporal information of the human pose sequence; (b) our method presents to predict multiple future poses in a non-recursive manner, which can efficiently prevent noise accumulation from early-stage prediction.

### 4.3 Qualitative results

To further show the performance of our proposed method, the qualitative results of our model are given in Fig. 3. In the figure, the poses of the left column are the history poses, the blue poses of the right column denote the predicted poses of our proposed method. From top to bottom, it is the motion sequence of actions ‘Walkingtogether’, ‘Walkingdog’, ‘Walking’, ‘Greeting’, ‘Directions’, ‘Discussion’, ‘Eating’, and ‘Walking’, respectively.

![Image of predicted poses](https://example.com/pose_images)

Taking action ‘walking’ as an example, (i) compared with the RNN-based baseline ‘Res[8]’, the errors of ‘80’, ‘160’, ‘320’, and ‘400 ms’ decrease by 12, 21, 27, and 31, respectively. Since our proposed CNN model is built with temporal convolution, incorporating our special organised input tensor of the skeletal sequence, we can model both the spatial and temporal information efficiently. ‘Res[8]’ built its model entirely based on GRU, which cannot capture the spatial structure of the human body efficiently. Therefore, compared with the RNN’s baseline [8], our model can significantly improve the predictive performance. (ii) Compared with the CNN-based baseline ‘Conv[11]’, the errors of ‘80’, ‘160’, ‘320’, and ‘400 ms’ decrease by 5, 12, 18, and 22, respectively. Our residual convolutional encoder–decoder model enables our network to capture the strong correlations of the joints of one limb by our special organisation of the input tensor. Also, the global structure information of the human body and long-term temporal information of the human motions can be captured by stacking multiple residual convolutional blocks to enlarge the receptive field from both spatial and temporal dimensions. However, ‘Conv[11]’ modelled the spatial and temporal information from the previous human motion sequence using a plain CNN, and captured the spatial correlations of joints from different limbs with a large convolutional kernel designed by manual, which cannot model the spatial structure information of the human body efficiently. So that, compared with the CNN’s baseline [10], our residual CNN’s framework can better model the spatio-temporal information of the human motion sequence. Moreover, our network predicts a window of future poses in a non-recursive manner, while ‘Conv[11]’ predicted multiple future poses recursively. So that the later predictive poses of our model can avoid the inference of the noise from the early predictive poses efficiently. This may be one of the reasons for our better performance at later timestamps such as ‘320’ and ‘400 ms’ than other baselines.

Moreover, there shows a similar performance on average. Also, compared with baselines, our proposed method achieves the best performance on average, especially for long-term prediction, which evidence the effectiveness of our proposed method again. Specifically, compared with ‘Res[8]’, the errors of ‘80’ and ‘160 ms’ decrease by 10 and 24, respectively, while the errors of ‘320’ and ‘400 ms’ decrease by 39 and 44, respectively; compared with ‘Conv[11]’, the errors of ‘160’, ‘320’, and ‘400 ms’ decrease by 5, 7, and 8, respectively, and the error of ‘80’ is slightly worse, but the error is very close.

In summary, compared with both RNN’s [8] baseline and CNN’s [10] baseline, our proposed method achieves superior predictive performance, which may benefit from two folds: (a) our residual convolutional encoder enables the network to capture the hierarchical representation of the input pose signals and the strong correlations of joints of one limb, which can better describe the spatial and temporal information of the human pose sequence; (b) our method presents to predict multiple future poses in a non-recursive manner, which can efficiently prevent noise accumulation from early-stage prediction.

### 4.3 Qualitative results

To further show the performance of our proposed method, the qualitative results of our model are given in Fig. 3. In the figure, the poses of the left column are the history poses, the blue poses of the right column denote the predicted poses of our proposed method. From top to bottom, it is the motion sequence of actions ‘Walkingtogether’, ‘Walkingdog’, ‘Walking’, ‘Greeting’, ‘Directions’, ‘Discussion’, ‘Eating’, and ‘Walking’, respectively.

![Image of predicted poses](https://example.com/pose_images)

| Action                        | Baseline | Ours   | Improvement |
|-------------------------------|----------|--------|-------------|
| Sitting down                  | 20       | 18     | 2            |
| Taking photo                  | 18       | 16     | 2            |
| Sitting                       | 20       | 18     | 2            |
| Walking together              | 18       | 16     | 2            |
| Waiting                       | 20       | 18     | 2            |
| Greeting                      | 20       | 18     | 2            |
| Directions                    | 20       | 18     | 2            |
| Discussion                    | 20       | 18     | 2            |
| Eating                        | 20       | 18     | 2            |
| Walking dog                   | 20       | 18     | 2            |

Bold values indicate the lowest errors.
larger errors in these limbs. Specifically, for action ‘Walkingtogether’, the main errors occur in the right leg and the right hand of the human body at an early stage, while the main errors exist in the head and the left leg of the human body at a later stage; for action ‘Discussion’, the errors of the right upper limb are relatively large at early prediction, while the errors of the left hand and feet are relatively large at later's prediction; for action ‘Walking’, the errors mainly exist in the two hands and two legs of the human body. Also, these phenomena are consistent with the above analysis.

Moreover, although the larger errors have been made in the early stage, the errors of the corresponding joints in the later stage did not increase sharply. Taking ‘Walkingtogether’ as an example, compared with the error of joint ‘right foot’ in the third predictive pose, the error of this joint in the fourth predictive poses is decreased. It shows that the previous predictive performance does not affect the later predictive results. This benefits from our model predicted multiple frames in a non-recursiveness manner, which can avoid the inference of the error from early predictive frames efficiently.

5 Conclusion

This paper has presented a new encoder-decoder framework, stacked residual blocks-based encoder-decoder, to predict future poses in a non-recursiveness manner, which captures the hierarchical spatial and temporal representation of the input signals and avoids the noise accumulation effectively. Experimental results outperform the state-of-the-art recursive models, which demonstrate the effectiveness of the proposed method.

6 Acknowledgments

This work was supported partly by the National Natural Science Foundation of China (Grant No. 61673192), the Fundamental Research Funds for the Central Universities (Grant No. 2020XD-A04-1, 2019RC27), and BUPT Excellent Ph.D. Students Foundation (CZX2019111).

7 References

[1] Bütepage, J., Kjellström, H., Krägic, D.: ‘Anticipating many futures: online human motion prediction and generation for human–robot interaction’. 2018 IEEE Int. Conf. on Robotics and Automation (ICRA), Prague, Czech Republic, 2018, pp. 1–9

[2] Mainprize, J., Berenson, D.: ‘Human-robot collaborative manipulation planning using early prediction of human motion’. 2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Tokyo, Japan, 2013, pp. 299–306

[3] Gu, L.-Y., Wang, Y.-X., Ramanan, D., et al.: ‘Few-shot human motion prediction via meta-learning’. Proc. European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 432–450

[4] Kong, Y., Fu, Y.: ‘Human action recognition and prediction: a survey’. arXiv preprint arXiv:1806.11230, 2018

[5] Zhou, J., Jia, X., Shen, L., et al.: ‘Improved softmax loss for deep learning-based face and expression recognition’, Cogn. Comput. Syst., 2019, 1, (4), pp. 97–102

[6] Fragiadaki, K., Levine, S., Felsen, P., et al.: ‘Recurrent network models for human dynamics’. Proc. IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 4346–4354

[7] Jain, A., Zamir, A.R., Savarese, S., et al.: ‘Structural-RNN: deep learning on spatio-temporal graphs’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 5308–5317

[8] Martinez, J., Black, M.J., Romero, J.: ‘On human motion prediction using recurrent neural networks’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017, pp. 610–617

[9] Chu, H.-K., Adeli, E., Wang, B., et al.: ‘Action-agnostic human pose forecasting’. 2019 IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, Hawaii, USA, 2019, pp. 1432–1432

[10] Gous, P., Song, J., Aksan, E., et al.: ‘Learning human motion models for long-term predictions’. 2017 Int. Conf. on 3D Vision (3DV), Qingdao, China, 2017, pp. 458–466

[11] Tang, Y., Ma, L., Liu, W., et al.: ‘Long-term human motion prediction by modeling motion context and enhancing motion dynamic’. arXiv preprint arXiv:1805.02513, 2018

[12] Guo, X., Chai, J.: ‘Human motion prediction via learning local structure representations and temporal dependencies’. Proc. AAAI Conf. on Artificial Intelligence, Honolulu, Hawaii, USA, 2019, vol. 33, pp. 2580–2587

[13] Liu, Z., Wu, S., Jin, S., et al.: ‘Towards natural and accurate future motion prediction of humans and animals’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 10004–10012

[14] Du, X., Vasudevan, R., Johnson-Roberson, M.: ‘Bio-LSTM: a biomechanically inspired recurrent neural network for 3D pedestrian pose and gait prediction’. IEEE Robot. Autom. Lett., 2019, 4, (2), pp. 1501–1508

[15] Gui, L.-Y., Wang, Y.-X., Liang, X., et al.: ‘Adversarial geometry-aware human motion prediction’. Proc. European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 786–803

[16] Kundu, J.N., Gor, M., Babu, R.V.: ‘BiHMP-GAN: bidirectional 3D human motion prediction GAN’. Proc. AAAI Conf. on Artificial Intelligence, Honolulu, Hawaii, USA, 2019, vol. 33, pp. 8553–8560

[17] Hernandez, A., Gall, J., Moreno-Noguer, F.: ‘Human motion prediction via spatio-temporal inference’. Proc. IEEE Int. Conf. on Computer Vision, Seoul, Korea, 2019, pp. 7134–7143

[18] Barsoum, E., Kender, J., Liu, Z.: ‘HP-GAN: probabilistic 3D human motion prediction via GAN’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 2018, pp. 1418–1427

[19] He, K., Zhang, X., Ren, S., et al.: ‘Deep residual learning for image recognition’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–777

[20] Ionescu, C., Papava, D., Olaru, V., et al.: ‘Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments’, IEEE Trans. Pattern Anal. Mach. Intell., 2014, 36, (7), pp. 1325–1339

[21] Ionescu, C., Li, F., Simoncini, C.: ‘Latent structured models for human pose estimation’. Proc. IEEE Int. Conf. on Computer Vision, Barcelona, Spain, 2011

[22] Mao, W., Liu, M., Salzmann, M., et al.: ‘Learning trajectory dependencies for human motion prediction’. Proc. IEEE Int. Conf. on Computer Vision, Seoul, Korea, 2019, pp. 4989–4997