基于双向门控循环单元的3D人体运动预测

桑海峰 陈紫珍*

(沈阳工业大学信息科学与工程学院 沈阳 110870)

摘 要：在机器视觉领域，预测人体运动对于及时的人机交互及人员跟踪等是非常有必要的。为了改善人机交互及人员跟踪等的性能，该文提出一种基于双向门控循环单元(GRU)的编-解码器模型(EBiGRU-D)来学习3D人体运动并给出一段时间内的运动预测。EBiGRU-D是一种深递归神经网络(RNN)，其中编码器是一个双向GRU(BiGRU)单元，解码器是一个单向GRU单元。BiGRU使原始数据从正反两个方向同时输入并进行编码，编成一个状态向量然后送入解码器进行解码。BiGRU将当前的输出与前后时刻的状态关联起来，使输出充分考虑了前后时刻的特征，从而使预测更加准确。在human3.6m数据集上的实验表明EBiGRU-D不仅极大地改善了3D人体运动预测的误差还大大地增加了准确预测的时间。

关键词：人体运动预测；3D动作识别；递归神经网络；深度学习

中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2019)09-2256-08

DOI: 10.11999/JEIT180978

3D Human Motion Prediction Based on Bi-directional Gated Recurrent Unit

SANG Haifeng CHEN Zizhen

(*School of Information Science & Engineering*, *Shenyang University of Technology*, *Shenyang* 110870, *China*)

Abstract: In the field of computer vision, predicting human motion is very necessary for timely human–computer interaction and personnel tracking. In order to improve the performance of human–computer interaction and personnel tracking, an encoder-decoder model called Bi–directional Gated Recurrent Unit Encoder–Decoder (EBiGRU–D) based on Gated Recurrent Unit (GRU) is proposed to learn 3D human motion and give a prediction of motion over a period of time. EBiGRU–D is a deep Recurrent Neural Network (RNN) in which the encoder is a Bidirectional GRU (BiGRU) unit and the decoder is a unidirectional GRU unit. BiGRU allows raw data to be simultaneously input from both the forward and reverse directions and then encoded into a state vector, which is then sent to the decoder for decoding. BiGRU associates the current output with the state of the front and rear time, so that the output fully considers the characteristics of the time before and after, so that the prediction is more accurate. Experimental results on the human3.6m dataset demonstrate that EBiGRU–D not only improves greatly the error of 3D human motion prediction but also increases greatly the time for accurate prediction.

Key words: Human motion prediction; 3D motion recognition; Recurrent Neural Network (RNN); Deep learning

1 引言

近年来，关于人体运动预测[1–3]的研究越来越受到人们的关注。人体运动预测作为机器人智能重要组成部分，可以使机器人对复杂的环境变化做出快速和高保真的反应。例如，机器人可以通过预测其周围主体的运动来毫不费力地避免路线碰撞等。如今，随着Kinect等MOCAP设备，以及姿态估计算法[4,5]等的发展，可以轻松准确地计算人体骨骼的序列。因此能够通过分析观察到的骨架序列来预测人体的运动，进一步帮助人体动作的分析与识别，身体姿势估计，甚至是人机交互。在过去几年中，基于视频及mocap数据的人体运动预测取得了重大进展。人体的运动表示[6–8]不仅需要动作分类，还需要运动预测和生成。而且骨架姿势和运动

收稿日期：2018-10-19；改回日期：2019-03-08；网络出版：2019-04-09
*通信作者： 陈紫珍 chenziz@126.com
基金项目：国家自然科学基金(61773105)，辽宁省自然科学基金(20170540675)，辽宁省教育厅科研项目(LQGD2017023)

Foundation Items: The National Natural Science Foundation of China (61773105), The Natural Science Foundation of Liaoning Province (20170540675), The Research Project of Liaoning Provincial Department of Education (LQGD2017023)

的一般表示可以在不同领域中用于不同目的。在计算机视觉中，适当的运动表示可以促进目标跟踪和识别。在机器人技术中，该表示可以为目标导向动作的意图推断和解释奠定基础。因此，这就要求一种足够有效的表示，使其可推广到新颖的运动中。此外，该表示还需要编码关节和肢体之间的相关性以及人体运动的时间结构。这种工作的目的是开发和研究骨骼人体运动数据的学习表示，使其可用于各种任务，而不仅是针对特定的运动模式。近年来，关于人体运动预测的方法越来越多也越来越完善。

在早期的工作中，Taylor等人[9]提出了一种具有2元隐变量的自回归受限玻尔兹曼机，用于人体运动预测。但他们的实验仅限于步行、慢跑和快跑运动。这种模型并不具有通用性。2014年，Fragkiadaki等人[10]介绍了一种编码器-递归-解码器网络，该网络通过长短期记忆(Long Short-Term Memory, LSTM)模型来学习人体运动的时间动态。他们设计了一个非线性变换来编码姿势特征并解码LSTM的输出得到预测值。历史信息在整个循环单元中传递以约束人体运动预测。但是记忆单元存储的信息有限，因此限制了准确预测的时长。2015年，Holden等人[11]使用单层卷积自动编码器来学习人体运动的低维流形。对于运动合成，学习特征和高级动作命令形成了前馈网络的输入，该前馈网络以重建期望的运动模式被训练。但这种流形学习的想法，使用了卷积层和池化层使层次结构更加复杂了。同年，Ashesh等人[12]以结构递归神经网络(Structure-Recurrent Neural Network, S-RNN)的形式引入了循环网络和人体结构层次的组合用于运动预测。通过构造由LSTM组成的节点和边缘的结构图，对个体肢体和整个身体的时间动态进行建模。该方法在没有低维表示的帮助下，还要针对每个运动训练单个模型。因此，该方法的计算比较复杂，成本较高。2016年，Martinez等人[13]通过对关节速度建模而不是直接估计身体姿势进一步扩展了Fragkiadaki等人的方案，并采用单线性层进行姿势特征编码，然后用隐藏状态解码得到预测值。该文发现0速度姿态在平均角度距离上的误差相对较小，证明了速度建模的效率。但是该模型随着运动时长增加时，误差依旧无明显改善。2018年，Tang等人[14]提出了改进高速单元(Modified Highway Unit, MHU)模型，通过探索运动背景和增强运动动力来预测长期人体运动。此外，他们还通过最小化了长矩阵运动预测的克矩阵(gram matrix)损失来增强运动动态。用于有效地消除关节静止并在给定运动环境的情况下估计下一个姿势。这种模型的确改善了长期运动的效果，但不是很明显。

本文提出了基于双向门控循环单元(Gated Recurrent Unit, GRU)的编-解码器模型(Bi–directional Gated Recurrent Unit Encoder–Decoder, EBiGRU-D)，这是一个基于时间序列数据的预测模型，用来学习在给定前几帧窗口的情况下预测许多未来的mocap帧。BiGRU让数据从正反两个方向同时输入，使每一时刻的信息都包含了前后时刻的序列信息，相当于网络在某个特定时刻的编码信息就多了，解码预测时，考虑的信息自然也就增多了，从而使预测的动作更加准确。本文的主要贡献有：(1)提出了一种无监督表示学习方案，用于改善人体运动预测的长期效果，本方案具有通用性，并不局限于一小部分动作；(2)提出的EBiGRU-D网络结构简单，易实现，计算简单，训练时间明显缩短。

2　EBiGRU-D预测原理

2.1　EBiGRU-D整体网络结构

本文所提EBiGRU-D预测模型见图1。模型大体上由2部分组成，BiGRU编码器及GRU解码器。从总结构图可以看出GRU是EBiGRU-D网络不可或缺的一部分。由于本文中编码和解码都需要用到GRU的计算，因此在这一部分，先介绍GRU的前向计算和反向计算原理[15,16]。BiGRU编码原理和GRU解码原理分别见2.2和2.3节。GRU是RNN的一种变体，RNN是深度学习中用于处理时序数据的关键技术，目前已在自然语言处理，语音识别，视频识别等领域取得重要突破，然而梯度消失现象制约着RNN的实际应用。GRU是目前广为使用的RNN变体，它通过门控机制很大程度上缓解了RNN的梯度消失问题。图2是GRU的结构图，图中的z_t和r_t分别为更新门和重置门。更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大说明前一时刻的状态信息带入越多。重置门用于控制忽略前一时刻的状态信息的程度，重置门的值越小说明忽略得越多。

GRU单元前向计算时首先通过上一个传输下来的状态$\boldsymbol{h}_{t-1}$和当前节点的输入x_t来获取2个门控状态。

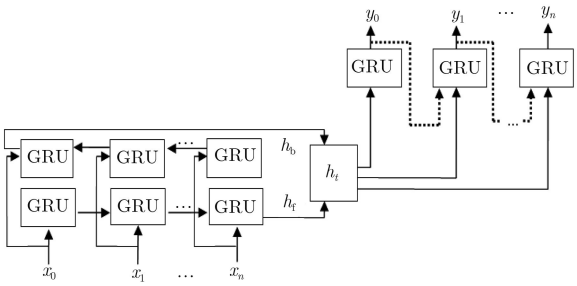

图 1 EBiGRU-D网络结构

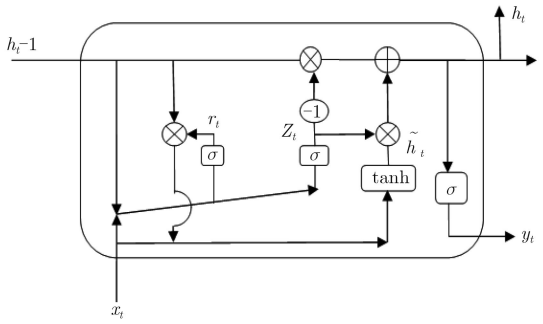

图 2 GRU内部结构图

$$\left.\begin{aligned} r_t &= \sigma\left(\boldsymbol{W}_r \cdot [\boldsymbol{h}_{t-1}, x_t]\right) \\ z_t &= \sigma\left(\boldsymbol{W}_z \cdot [\boldsymbol{h}_{t-1}, x_t]\right) \end{aligned}\right\} \tag{1}$$

其中，$\boldsymbol{W}_r$, $\boldsymbol{W}_z$表示权重矩阵。

得到门控信号之后，首先使用重置门控来得到“重置”之后的数据$r_t \times \boldsymbol{h}_{t-1}$，再将$r_t \times \boldsymbol{h}_{t-1}$与输入$x_t$进行拼接，再通过tanh激活函数将数据放缩到–1～1的范围内。

$$\tilde{\boldsymbol{h}}_t = \tanh \boldsymbol{W}_{\tilde{h}} \cdot [r_t \times \boldsymbol{h}_{t-1}, x_t] \tag{2}$$

其中，$\boldsymbol{W}_{\tilde{h}}$表示权重矩阵，$\tilde{\boldsymbol{h}}_t$主要是包含了当前输入的$x_t$数据。有针对性地将$\tilde{\boldsymbol{h}}_t$添加到当前的隐藏状态，这一步相当于“记忆了当前时刻的状态”。

GRU最关键的一个步骤，即“更新记忆”阶段。当前隐藏状态$\boldsymbol{h}_t$更新表达式如式(3)

$$\boldsymbol{h}_t = 1 - z_t \times \boldsymbol{h}_{t-1} + z_t \times \tilde{\boldsymbol{h}}_t \tag{3}$$

更新门控z_t的范围是0～1，门控信号越接近1，代表“记忆”的数据越多，越接近0代表“遗忘”的越多。这也是GRU的优点表现之一，GRU使用同一个门控z_t就可以同时进行遗忘和选择记忆，而LSTM则需要使用多个门控来达到这种效果，相比之下，GRU的计算就更简单了。本文使用的GRU单元进行编-解码时，仅需要获取隐藏状态$\boldsymbol{h}_t$的输出即可。

GRU的后向计算其实就是训练$\boldsymbol{W}_r$, $\boldsymbol{W}_z$, $\boldsymbol{W}_h$这些权重参数，这3个权重都是拼接的，所以在学习时需要分割出来，即

$$\left.\begin{aligned} \boldsymbol{W}_r &= \boldsymbol{W}_{rx} + \boldsymbol{W}_{rh} \\ \boldsymbol{W}_z &= \boldsymbol{W}_{zx} + \boldsymbol{W}_{zh} \\ \boldsymbol{W}_{\tilde{h}} &= \boldsymbol{W}_{\tilde{h}x} + \boldsymbol{W}_{\tilde{h}h} \end{aligned}\right\} \tag{4}$$

设某时刻的损失函数为$E_t = (y_d - y_{t_0})^2/2$，则某样本的损失为$E = \sum_{t=1}^{T} E_t$。$y_d$表示$t$时刻的真实值。假设当前时刻$t$的误差项为$\delta_t = \partial E/\partial \boldsymbol{h}_t$，那么误差沿着时间反向传递则需要计算$t-1$时刻的误差项$\delta_{t-1}$，则$\delta_{t-1} = \partial E/\partial \boldsymbol{h}_{t-1} = \partial E/\partial \boldsymbol{h}_t \partial \boldsymbol{h}_t/\partial \boldsymbol{h}_{t-1} = \delta_t \partial \boldsymbol{h}_t/\partial \boldsymbol{h}_{t-1}$。$\boldsymbol{h}_t$可看成是一个复合函数，由$r_t$, z_t, $\tilde{\boldsymbol{h}}_t$组成。根据全导数公式可计算出δ_{t-1}

$$\delta_{t-1} = \delta_{r,t}\boldsymbol{W}_{rh} + \delta_{z,t}\boldsymbol{W}_{zh} + \delta_{\tilde{h},t}\boldsymbol{W}_{\tilde{h}h} \tag{5}$$

接着对某时刻t的所有权重求偏导，即

$$\left.\begin{aligned} \frac{\partial E}{\partial \boldsymbol{W}_{zx}} &= \delta_{z,x} x_t \\ \frac{\partial E}{\partial \boldsymbol{W}_{zh}} &= \delta_{z,t} \boldsymbol{h}_{t-1} \\ \frac{\partial E}{\partial \boldsymbol{W}_{\tilde{h}x}} &= \delta_t x_t \\ \frac{\partial E}{\partial \boldsymbol{W}_{\tilde{h}h}} &= \delta_t (r_t \cdot \boldsymbol{h}_{t-1}) \\ \frac{\partial E}{\partial \boldsymbol{W}_{rx}} &= \delta_{r,t} x_t \\ \frac{\partial E}{\partial \boldsymbol{W}_{rh}} &= \delta_{r,t} \boldsymbol{h}_{t-1} \end{aligned}\right\} \tag{6}$$

其中，GRU各部分误差项的求解如式(7)

$$\left.\begin{aligned} \delta_{\boldsymbol{h},t} &= \delta_{y,t}\boldsymbol{W}_o + \delta_{z,t+1}\boldsymbol{W}_{zh} + \delta_{t+1}\boldsymbol{W}_{\tilde{h}h} \cdot r_{t+1} \\ &\quad + \delta_{\boldsymbol{h},t+1}\boldsymbol{W}_{rh} + \delta_{\boldsymbol{h},t+1} \cdot (1 - z_{t+1}) \\ \delta_{z,t} &= \delta_{\boldsymbol{h},t} \cdot \left(\tilde{\boldsymbol{h}}_t - \boldsymbol{h}_{t-1}\right) \cdot \sigma' \\ \delta_t &= \delta_{\boldsymbol{h},t} \cdot z_t \cdot \varphi' \\ \delta_{r,t} &= \boldsymbol{h}_{t-1} \cdot [(\delta_{\boldsymbol{h},t} \cdot z_t \cdot \varphi')\, \boldsymbol{W}_{\tilde{h}h}] \cdot \sigma' \end{aligned}\right\} \tag{7}$$

其中，σ'和φ'分别表示sigmoid函数和tanh函数的导数，$\delta_{\boldsymbol{h},t}$是对$\boldsymbol{h}_t$函数求偏导的结果，$\delta_{z,t}$是对$z_t$函数求偏导的结果，$\delta_{r,t}$是对$r_t$函数求偏导的结果。

对于整个样本，它的误差是所有时刻的误差之和，与上个时刻相关的权重的梯度等于所有时刻的梯度之和，其它权重则不必累加，最终得到

$$\left.\begin{aligned} \frac{\partial E}{\partial \boldsymbol{W}_{zh}} &= \sum_{j=1}^{t} \delta_{z,j} \boldsymbol{h}_{j-1} \\ \frac{\partial E}{\partial \boldsymbol{W}_{\tilde{h}h}} &= \sum_{j=1}^{t} \delta_j (r_j \cdot \boldsymbol{h}_{j-1}) \\ \frac{\partial E}{\partial \boldsymbol{W}_{rh}} &= \sum_{j=1}^{t} \delta_{r,j} \boldsymbol{h}_{j-1} \\ \frac{\partial E}{\partial \boldsymbol{W}_{\tilde{h}x}} &= \delta_t x_t \\ \frac{\partial E}{\partial \boldsymbol{W}_{zx}} &= \delta_{z,x} x_t \\ \frac{\partial E}{\partial \boldsymbol{W}_{rx}} &= \delta_{r,t} x_t \end{aligned}\right\} \tag{8}$$

有了各个参数的偏导数之后，就可以用梯度下降法来更新各参数了，更新公式如式(9)

$$\left.\begin{aligned}
\boldsymbol{W}_{zx} &= \boldsymbol{W}_{zx} - \alpha\delta_{z,x}x_t \\
\boldsymbol{W}_{z\boldsymbol{h}} &= \boldsymbol{W}_{z\boldsymbol{h}} - \alpha\sum_{j=1}^{t}\delta_{z,j}\boldsymbol{h}_{j-1} \\
\boldsymbol{W}_{\tilde{\boldsymbol{h}}x} &= \boldsymbol{W}_{\tilde{\boldsymbol{h}}x} - \alpha\delta_t x_t \\
\boldsymbol{W}_{rx} &= \boldsymbol{W}_{rx} - \alpha\sum_{j=1}^{t}\delta_j\left(r_j\cdot\boldsymbol{h}_{j-1}\right) \\
\boldsymbol{W}_{\tilde{\boldsymbol{h}}x} &= \boldsymbol{W}_{\tilde{\boldsymbol{h}}x} - \alpha\delta_{r,t}x_t \\
\boldsymbol{W}_{r\boldsymbol{h}} &= \boldsymbol{W}_{r\boldsymbol{h}} - \alpha\sum_{j=1}^{t}\delta_{r,j}\boldsymbol{h}_{j-1}
\end{aligned}\right\} \tag{9}$$

其中，α为学习率。本文实验所选的α值为0.001。

2.2 BiGRU的编码原理

编码器一般用于读取源数据，然后产出一个在连续空间中的特征表示。本文编码器的作用是把一个不定长的输入序列转换成一个定长的状态向量$\boldsymbol{C}$，该状态向量包含了输入序列的信息。本文将BiGRU单元作为编码器，BiGRU的基本思想是将每个训练序列向前和向后呈现给两个单独的隐藏层，这两个层都连接到相同的输出层。因此输出层就具有了输入序列中每个点的完整过去和未来的信息，从而不会出现替换相关目标输入的情况。本文用到的BiGRU的结构见图3。一个正常的递归网络是包含输入层、隐藏层和输出层的，但本文使用BiGRU网络是为了对输入数据进行记忆，然后将记忆信息传递给解码器，才会得到预测值，因此本文的BiGRU只用到了输入层和隐藏层。编码器部分最终的目的就是获得储存了输入信息的最终隐藏状态$\boldsymbol{h}_t$。

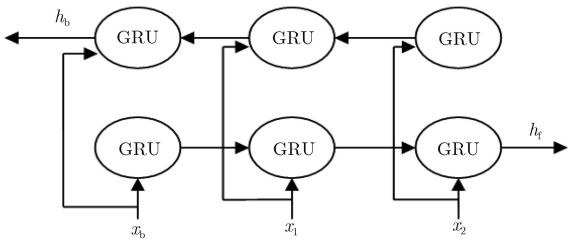

图 3 BiGRU部分结构图

BiGRU的前向计算跟单向GRU一样，但是双向GRU的输入序列对于两个隐含层是相反方向的，而且输出层直到两个隐含层处理完所有的全部输入序列才更新。BiGRU的后向计算也与GRU相似，所有的输出层δ项首先被计算，然后再返回给两个不同方向的隐含层。

在得到正向隐藏状态输出和反向隐藏状态输出后，将两者进行拼接即可得到编码器的隐藏状态$\boldsymbol{h}_t$

$$\boldsymbol{h}_t = \left[\boldsymbol{h}_f^t, \boldsymbol{h}_b^t\right] \tag{10}$$

其中$[\cdot,\cdot]$表示两个向量拼接，$\boldsymbol{h}_t$就是最终需要的编码向量$\boldsymbol{C}$。本文默认初始状态为0。得到$\boldsymbol{C}$之后就可将$\boldsymbol{C}$送入解码器进行解码，得到预测值了。

2.3 GRU的解码原理

解码阶段可以看做编码的逆过程。在这个阶段，本文根据给定的语义向量$\boldsymbol{C}$和之前已经生成的输出序列$y_0y_1\cdots y_n$来预测下一个输出值。即

$$y_t = g\left(y_{t-1}, s_t, \boldsymbol{C}\right) \tag{11}$$

其中，s_t是输出GRU中的隐藏层，$\boldsymbol{C}$是编码器输出的状态向量，y_{t-1}表示上个时间段的输出，用来作为这个时间段的输入。g是一个非线性变换，本文选取的是softmax函数。本文默认解码器初始状态为0。

3 实验及分析

本文在human3.6m数据集上进行了实验，并将结果与近期方法进行了对比。本文评估指标有2个：(1)定量预测误差，测量的是关节角度的均方损失，(2)较长时间范围的定性运动合成。human3.6m数据集是目前最大的视频姿态数据集。它由4台静态摄像机录制，由7名不同的专业演员执行的15个活动场景组成。对于每个活动场景，由于主题和相机视点不同，因此有2个视频序列，每个视频序列在3000～5000帧之间。每个活动场景都有丰富的手势，姿态变化等有趣的子动作。例如，步行活动包括牵手，背负重物，双手放在口袋里，四处张望等等。使用Vicon动作捕捉系统记录动作，追踪演员身体关节上的标记，并提供高品质的3D身体关节位置。通过使用已知的照相机校准和视点将3D位置投影到图像平面上来获得2D身体关节位置。对于本文所有的实验，科目5是测试科目，其他科目作为训练科目。训练时，学习率设置为0.001，批量大小为32，并将梯度限制在最大L2范数为5的范围内，每秒输入的帧数是25帧。

实验软件环境操作系统为Ubuntu 16.04，深度学习软件框架为TensorFlow。实验硬件环境是在配有8 G显存的GTX 1080图形处理器、Intel Core i7–7700K中央处理器和8 G内存的台式电脑上进行。

3.1 短期预测实验及分析

由于近期的研究方法预测的时间一般在1 s内，为了公平对比，因此本文第1个实验就是预测1 s内的动作，表1显示了本文所提方法即EBiGRU-D网络在human3.6m数据集上的测试结果与近期方法结果的对比。其中，近期方法的数据来源于他们所公开发表的数据。

从表1中数据可以看出本文所提EBiGRU-D网络在1 s内关于walking, greeting, walkingdog和

表 1 human3.6m数据集下1 s内各模型预测误差的对比(ms)

预测时间(ms)	80	160	320	400	560	640	720	1000
Walking								
ERD[10]	0.77	0.90	1.12	1.25	1.44	1.45	1.46	1.49
LSTM-3LR[10]	0.73	0.81	1.05	1.18	1.34	1.36	1.37	1.36
Res-GRU[13]	0.39	0.68	0.99	1.15	1.35	1.37	1.37	1.32
MHU[14]	0.32	0.53	0.69	0.77	0.90	0.94	0.97	1.06
EBiGRU-D	0.31	0.31	0.33	0.35	0.35	0.36	0.36	0.37
Greeting								
ERD[10]	0.85	1.09	1.45	1.64	1.93	1.89	1.92	1.98
LSTM-3LR[10]	0.80	0.99	1.37	1.54	1.81	1.76	1.79	1.85
Res-GRU[13]	0.52	0.86	1.30	1.47	1.78	1.75	1.82	1.96
MHU[14]	0.54	0.87	1.27	1.45	1.75	1.71	1.74	1.87
EBiGRU-D	0.48	0.44	0.49	0.49	0.52	0.51	0.52	0.49
Walkingdog								
ERD[10]	0.91	1.07	1.39	1.53	1.81	1.85	1.90	2.03
LSTM-3LR[10]	0.80	0.99	1.37	1.54	1.81	1.76	1.79	2.00
Res-GRU[13]	0.56	0.95	1.33	1.48	1.78	1.81	1.88	1.96
MHU[14]	0.56	0.88	1.21	1.37	1.67	1.72	1.81	1.90
EBiGRU-D	0.51	0.64	0.61	0.62	0.62	0.59	0.61	0.60
Discussion								
ERD[10]	0.76	0.96	1.17	1.24	1.57	1.70	1.84	2.04
LSTM-3LR[10]	0.71	0.84	1.02	1.11	1.49	1.62	1.76	1.99
Res-GRU[13]	0.31	0.69	1.03	1.12	1.52	1.61	1.70	1.87
MHU[14]	0.31	0.67	0.93	1.00	1.37	1.56	1.66	1.88
EBiGRU-D	0.33	0.44	0.50	0.45	0.48	1.51	0.50	0.49

discussion这4个动作的预测误差较近期的方法有一定地提升。EBiGRU-D网络在80 ms处的预测误差值与MHU网络及Res-GRU网络的预测误差值相差不多，但EBiGRU-D网络的预测值随着时间的增加波动较小。而这些近期方法大多对前几帧的预测比较准确，对时间较长的预测并没有很好的预测效果。

图4显示了1 s内本文方法与Res-GRU[13]方法关于walking动作的定性运动合成结果的对比，图中首尾帧分别是预测的第1帧和最后1帧，中间是按时间每隔1帧选取的。由于其它近期方法的实验并没有公开进行，因此本实验无法进行定性运动合成的对比。为了与Res-GRU网络进行更公平地比较，本文所有的实验输入设置都是50帧，与Re-GRU网络测试的输入一致。

从图4可看出Res-GRU[13]网络预测的动作前几帧与真实动作相似度很高，但随着时间加长，身体越来越弯曲，而且真实动作的左手有明显的摆动幅度，但Res-GRU预测的结果左手摆动幅度明显跟不上。本文提出的EBiGRU-D网络预测的动作无论在身体直立程度还是手的摆动幅度方面较Res-GRU网络效果有一定的提升，与真实动作有较高的相似度。

为了进一步验证EBiGRU-D网络针对复杂动作依然有效，本文对human3.6m数据集中的复杂动作discussion也进行了定性运动合成并将结果与Res-GRU[13]网络预测的结果进行了对比，如图5所示。

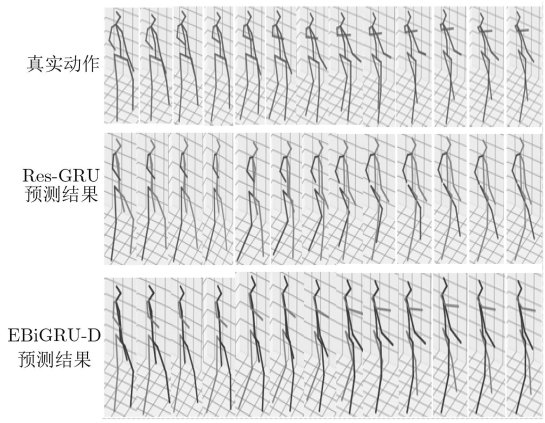

图 4 1 s内关于walking动作预测性能对比

从图5中可看出，Res-GRU[13]网络预测的前几帧与真实动作也较为接近，但是往后，身体明显地向前倒，而且幅度很大，双手的摆动幅度一开始也与真实动作相似，但随着时间的增加，双手与真实动作差距明显拉开。相比之下，本文所提EBiGRU-D网络预测的动作较为接近真实动作，身体也有向前倒，但幅度不是很明显，双手的摆动幅度虽然没有完全跟上真实动作，但相比Res-GRU网络，改善的也比较明显。

3.2 长期预测实验及分析

为了验证EBiGRU-D网络对于较长期的预测也有一定的效果，本节将对进行human3.6m数据集中的walking, greeting, walkingdog和discussion这4个动作进行长达2 s的预测。为了更公平地体现网络的性能，这一节中的输入依旧设置为50帧。除了输出变为50帧之外，其它实验设置与上一节一致。表2显示了2 s内在human3.6m数据集上EBiGRU-D网络的测试结果与Res-GRU网络的测试结果的对比。其中，Res-GRU网络的测试结果是根据文献[14]中的实验设置复现的结果。

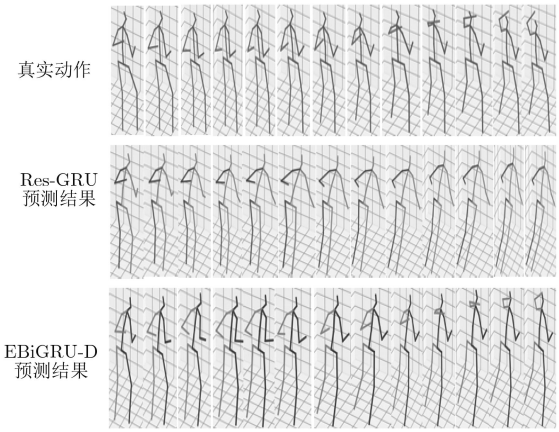

图 5 1 s内关于discussion动作的预测性能的对比

从表2中数据可以看出本文所提EBiGRU-D网络在2 s内关于walking, greeting, walkingdog和discussion这4个动作的预测误差较Res-GRU方法有明显的改善。Res-GRU方法的预测误差几乎都会随着时间加长，越来越大。而本文所提EBiGRU-D网络的预测值随着时间加长，变化并不明显，保持了比较低的误差值。

图6显示了2 s内关于walking动作的定性运动合成，图中首尾帧分别是预测的第1帧和最后1帧，中间是按时间每隔5帧的方式选取的。从图中可以看出Res-GRU网络预测的身体有点弯曲，左手摆动幅度也明显与真实动作有差距。EBiGRU-D网络预测的身体不弯曲，但有点向前倾。手的摆动程度虽然没有达到真实动作的高度，但相比之下，改善得也比较明显。

Walking属于一个比较简单的、容易执行的动作，光从walking的预测结果并不太能证明本文模型对长期预测的有效性。因此，为了验证EBiGRU-D网络对于复杂的动作也能进行较长期的预测，本文对较为复杂的动作walkingdog和discussion进行了长达2 s的定性运动合成。效果如图7所示，选取方式与图6一致。

从图7可以看出，EBiGRU-D网络预测的动作与真实动作具有较高的相似性。关于walkingdog动作的预测，Res-GRU网络的预测动作除了第1帧比较接近真实动作外，其它帧与真实动作差距较大，特别是双腿，整个过程双腿几乎是飘在半空的。相比之下，EBiGRU-D网络预测的动作就比较接近真

表 2　human3.6m数据集下2 s内EBiGRU-D网络和Res-GRU网络预测误差的对比(ms)

预测时间(ms)	80	320	560	720	1000	1080	1320	1560	1720	2000
Walking										
Res-GRU[13]	0.42	0.89	1.02	1.16	1.37	1.39	1.46	1.59	1.65	1.89
EBiGRU-D	0.36	0.35	0.36	0.39	0.41	0.41	0.44	0.47	0.48	0.48
Greeting										
Res-GRU[13]	0.65	0.89	1.21	1.35	1.56	1.77	1.85	2.02	2.16	2.22
EBiGRU-D	0.45	0.46	0.50	0.51	0.55	0.54	0.56	0.56	0.55	0.56
Walkingdog										
Res-GRU[13]	0.66	1.20	1.73	1.95	2.20	2.27	2.34	2.41	2.51	2.52
EBiGRU-D	0.49	0.58	0.60	0.60	0.61	0.60	0.59	0.60	0.61	0.61
Discussion										
Res-GRU[13]	0.89	1.23	1.56	1.69	1.85	2.01	2.12	2.32	2.49	2.56
EBiGRU-D	0.42	0.43	0.43	0.45	0.49	0.50	0.55	0.55	0.54	0.56

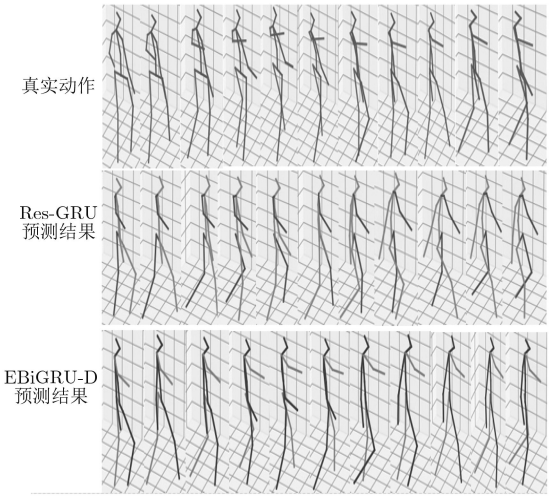

图 6 2 s内关于walking动作预测性能的对比

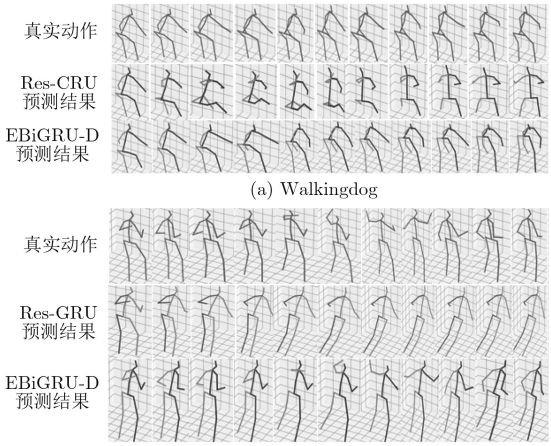

图 7 2 s内关于复杂动作的EBiGRU-D网络和Res-GRU网络性能的对比

实动作了，但还是存在一定的缺陷的，比如双腿的弯曲角度问题，真实动作有明显的屈膝，而EBiGRU-D网络的预测动作双膝弯曲不明显。关于discussion动作的预测，Res-GRU网络的预测动作前几帧还是比较接近真实动作的，但时间越往后，预测的动作几乎定格了，而且整个身体还往前倾斜。而EBiGRU-D网络的预测动作在整个过程中，摆动幅度都与真实动作接近，但也有不足的是，预测的动作有点向前倾斜。

关于本文实验部分更详细更直观结果请看https://pan.baidu.com/s/1StsnrScHJz_c9uYkgnv0XQ。

3.3 实验训练时间的对比

本文采用的GRU网络是LSTM网络的变形，LSTM有3个门控单元，而GRU只有2个，因此计算上相对LSTM来说更简单了。图8显示了EBiGRU-D网络和Res-GRU网络关于walking, greeting, walkingdog和discussion这4个动作进行2 s预测实验时的训练时间对比。从图8中可以看出，在相同的实验条件下，EBiGRU-D网络的训练时间比Res-GRU网络要低得多。EBiGRU-D的训练时间在10 min左右，而Res-GRU的训练时间在20 min以上。

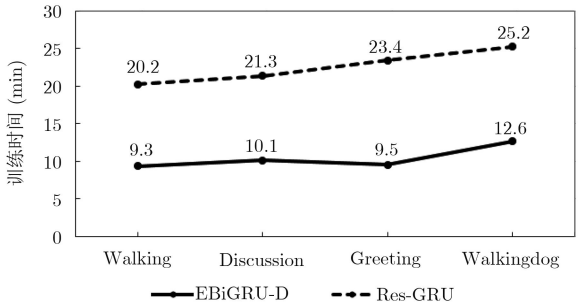

图 8 训练时间对比

4 结束语

本文主要提出了一个EBiGRU-D网络用于动作预测，通过融合双向门控循环单元加强了对输入数据的记忆及存储，最后通过简单的单向门控循环单元对存储下来的数据进行处理及解码得到预测值。本文提出的EBiGRU-D网络是一种无监督学习方式，实验已经证明了EBiGRU-D网络具有通用性，并不局限于一小部分动作。而且EBiGRU-D网络结构较为简单，并不需要较为昂贵的设备支持，最难得的是训练时间较短，对复杂的动作进行预测时，训练时间也仅需10 min左右。目前的预测方法一般都是针对单人且是在简单的环境下进行的，而且应用方面不是很广泛。下一步计划将针对多人或复杂的环境下进行预测，并将其及时应用于人机交互及人员跟踪方向。

参 考 文 献

[1] FOKA A F and TRAHANIAS P E. Probabilistic autonomous robot navigation in dynamic environments with human motion prediction[J]. *International Journal of Social Robotics*, 2010, 2(1): 79–94. doi: 10.1007/s12369-009-0037-z.

[2] MAINPRICE J and BERENSON D. Human–robot collaborative manipulation planning using early prediction of human motion[C]. 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2013: 299–306.

[3] BÜTEPAGE J, BLACK M J, KRAGIC D, *et al.* Deep representation learning for human motion prediction and classification[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1591–1599.

[4] TEKIN B, MÁRQUEZ–NEILA P, SALZMANN M, *et al.* Learning to fuse 2D and 3D image cues for monocular body pose estimation[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 3961–3970.

[5] YASIN H, IQBAL U, KRÜGER B, *et al.* A dual–source approach for 3D pose estimation from a single image[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4948–4956.

[6] 肖俊, 庄越挺, 吴飞. 三维人体运动特征可视化与交互式运动分割[J]. 软件学报, 2008, 19(8): 1995–2003.
XIAO Jun, ZHUANG Yueting, and WU Fei. Feature visualization and interactive segmentation of 3D human motion[J]. *Journal of Software*, 2008, 19(8): 1995–2003.

[7] 潘红, 肖俊, 吴飞, 等. 基于关键帧的三维人体运动检索[J]. 计算机辅助设计与图形学学报, 2009, 21(2): 214–222.
PAN Hong, XIAO Jun, WU Fei, *et al.* 3D human motion retrieval based on key-frames[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2009, 21(2): 214–222.

[8] LI Rui, LIU Zhenyu, and TAN Jianrong. Human motion segmentation using collaborative representations of 3D skeletal sequences[J]. *IET Computer Vision*, 2018, 12(4): 434–442. doi: 10.1049/iet-cvi.2016.0385.

[9] TAYLOR G W, HINTON G E, and ROWEIS S. Modeling human motion using binary latent variables[C]. The 19th International Conference on Neural Information Processing Systems, Hong Kong, China, 2006: 1345–1352.

[10] FRAGKIADAKI K, LEVINE S, FELSEN P, *et al.* Recurrent network models for human dynamics[C]. The IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4346–4354.

[11] HOLDEN D, SAITO J, and KOMURA T. A deep learning framework for character motion synthesis and editing[J]. *ACM Transactions on Graphics*, 2016, 35(4): 1–11.

[12] ASHESH J, ZAMIR A R, SAVARESE S, *et al.* Structural-RNN: Deep learning on spatio–temporal graphs[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 5308–5317.

[13] MARTINEZ J, BLACK M J, and ROMERO J. On human motion prediction using recurrent neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 4674–4683.

[14] TANG Yongyi, MA Lin, LIU Wei, *et al.* Long–term human motion prediction by modeling motion context and enhancing motion dynamic[J/OL]. arXiv: 1805.02513. http://arxiv.org/abs/1805.02513, 2018.

[15] ZHANG Yachao, LIU Kaipei, QIN Liang, *et al.* Deterministic and probabilistic interval prediction for short&-term wind power generation based on variational mode decomposition and machine learning methods[J]. *Energy Conversion and Management*, 2016, 112: 208–219. doi: 10.1016/j.enconman.2016.01.023.

[16] CHO K, VAN MERRIENBOER B, GULCEHRE C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation[J/OL]. arXiv: 1406.1078, 2014.

桑海峰：男，1978年生，教授，博士，研究方向为视觉检测技术与图像处理，人工智能.

陈紫珍：女，1994年生，硕士生，研究方向为计算机视觉与图像处理，人工智能.