计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-3039-13
doi: 10.3778/j.issn.1673-9418.2208039

多尺度全局自适应注意力图神经网络

苟茹茹[1],杨文柱[1,2+],罗梓菲[1],原云峰[1]
1. 河北大学 网络空间安全与计算机学院,河北 保定 071002
2. 河北大学 河北省机器视觉工程研究中心,河北 保定 071002
+ 通信作者 E-mail: wenzhuyang@hbu.edu.cn

摘 要:针对动态多尺度图神经网络的编解码网络中存在的身体部位内部关节点间关联度不高和感受野受限制导致运动预测误差偏高的问题,提出了一种用于人体运动预测的多尺度全局自适应注意力图神经网络,降低运动预测误差。提出了一种划分骨架关节点的多距离分区策略,用于提高身体部位关节点信息在时间和空间上的关联程度;提出了全局自适应注意力时空卷积神经网络,以动态地加强网络对某一动作有贡献的时空关节点的关注度;将上述两处改进集成到图卷积神经网络门控循环单元中,以增强解码网络的状态传播性能,并降低预测误差。实验表明,与最新方法相比,该方法在Human 3.6M、CMU Mocap和3DPW数据集上的预测误差都有所下降。

关键词:运动预测;多距离分区策略;全局自适应注意力;时空图卷积神经网络;门控循环单元
文献标志码:A **中图分类号:**TP391.41

Multiscale Global Adaptive Attention Graph Neural Network

GOU Ruru[1], YANG Wenzhu[1,2+], LUO Zifei[1], YUAN Yunfeng[1]
1. School of Cyber Security and Computer, Hebei University, Baoding, Hebei 071002, China
2. Hebei Machine Vision Engineering Research Center, Hebei University, Baoding, Hebei 071002, China

Abstract: Dynamic multiscale graph neural networks have high motion prediction errors due to the low correlation between the internal joints of body parts and the limited perceptual fields. A multiscale global adaptive attention graph neural network for human motion prediction is proposed to reduce motion prediction errors. Firstly, a multi-distance partitioning strategy for dividing skeleton joint is proposed to improve the degree of temporal and spatial correlation of body joint information. Secondly, a global adaptive attention spatial temporal graph convolutional network is designed to dynamically enhance the network's attention to the spatial temporal joints contributing to a motion in combination with global adaptive attention. Finally, this paper integrates the above two improvements into the graph convolutional neural network gate recurrent unit to enhance the state propagation performance of the decoding network and reduce prediction errors. Experimental results show that the prediction error of the proposed method is decreased on Human 3.6M dataset, CMU Mocap dataset and 3DPW dataset compared with state-of-the-art methods.

Key words: motion prediction; multi-distance partitioning strategy; global adaptive attention; spatial temporal graph convolution neural network; gated recurrent unit

人体运动预测已成为当今计算机视觉的研究热点之一。在众多领域都有广泛的应用,例如:自动驾驶、智能视频监控、智能医疗监护、人机交互和人体跟踪[1]等。以往的方法中大多使用隐马尔可夫链、受限玻尔兹曼机、随机森林和高斯过程动力学模型等,在简单的周期运动中取得了超出预想的效果,但是

收稿日期:2022-08-11 **修回日期:**2022-10-18

对于复杂动作的预测结果却不尽如人意。由于人体运动具有非周期性和随机性的特点,准确预测未来运动姿势仍是一项具有挑战性的任务。

随着深度学习方法的发展,循环神经网络[2]、卷积神经网络[3-4]和生成对抗神经网络[5-6]都在解决这一挑战上取得了重大的突破。但是这些方法的预测精确度都受卷积滤波器大小和逐帧预测稳定性的影响,且都忽略了运动中身体关节的时间相关性。基于关节点的图卷积神经网络[7]能够很好地捕捉运动中身体关节的时间相关性[8]。因此,基于骨架的图卷积神经网络已广泛应用于运动预测和其他各个领域,且取得了良好的效果。Mao等人[9]设计了一个完全联通的图卷积,以自适应地学习运动预测所需要的连通信息,并应用离散余弦变换[10-13]构建了跨身体关节的图形处理时间信息,从而实现了对成对关系的建模。但是这样的图表仍然不足以反映身体关节组件之间的联系。Wang等人[14]设计了新型的深度学习网络来模拟时空方差,并通过预定义结构构建了身体关节特征,用于表示固定的身体部位,但是该模型没有利用运动的协同关系。例如,“招手”的动作往往是基于抽象的手臂和手的协同运动来预测的,而不是手臂和手指的详细位置。为解决这些问题,Li等人[15]通过人体姿势的自然层次结构,借助时空图卷积神经网络与多尺度[16]联合构建编码器提取丰富的运动特征,通过基于可训练图卷积神经网络的门控循环单元[17]构建解码器结构来生成运动的未来姿势。但是该模型存在以下三个问题:(1)编解码网络中的时空图卷积神经网络使用的分区策略不利于提取身体部位内部关节点之间的关联关系;(2)编码网络中使用的时空图卷积网络局限于每个节点的共享变换矩阵,不利于全局特征的学习,不利于网络排除不相关关节点,且动态地关注对动作贡献度高的关节点;(3)解码网络中的可训练图卷积神经网络存在同样的问题,导致网络的预测误差较大。

基于此,本文对动态多尺度图卷积模型(dynamic multiscale graph neural networks,DMGNN)进行了改进,主要贡献包括:(1)提出了多距离分区策略(multi-distance partitioning strategy,MD),该分区策略加强了身体部位内部关节点的相对位置之间的联系,有利于提高身体关节点信息在空间和时间上的联系;(2)通过时空图卷积神经网络的时空块和非局部网络,组成全局自适应注意力时空图卷积神经网络(global adaptive attention spatial temporal graph convolutional network,GaST-GCN)模块,动态地捕捉骨架中高贡献关节点的时空的全局和远程关系,以解决接受域有限及无关关节干扰的问题;(3)在图卷积门控单元的图卷积中使用多距离分区策略以及全局自适应注意力,组成多距离全局自适应注意力图卷积门控循环单元(multi-distance partitioned global adaptive attention graph convolution gate recurrent unit,MGG-GRU),这样既可以保证关节点间的局部联系,又可以提高全局中高贡献关节点的关注度,从而增强解码网络的状态传播性能。

1 相关工作

1.1 ST-GCN算法

ST-GCN(spatial temporal graph convolution networks)[18]算法不同于使用递归神经网络和临时卷积神经网络构建的端到端的动作识别模型[19-21],是一种基于关节点的动作识别算法。ST-GCN首次将图卷积神经网络应用于基于骨架的人体动作识别中。在此基础上通过人体自然连接和相同关节的跨连续时间连接构建了骨架序列的时空图,从而加入了对识别人体行为非常重要的关节点之间的时空关系这一因素,使得信息可以沿着图和时间维度进行整合。

ST-GCN通过三种分区策略设计空间卷积核。如图1所示,其中(a)为输入骨架。(b)为单标签分区,将整个邻域分为一个子集。(c)为距离分区,通过节点之间的距离设置分区。选取 $K=1$,将邻域分为两个

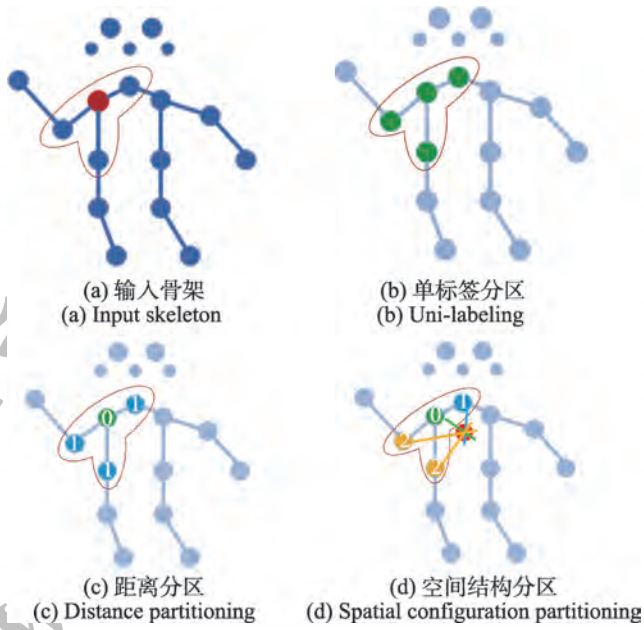

图1 分区策略

Fig.1 Partitioning strategy

子集，分别是距离为0的子集和距离为1的子集。得到两个不同的权重向量，能够对关节点的局部差异性进行建模。(d)为空间结构分区，通过人体运动特性确定人体关节点中的重心节点(一个框架中所有关节点的平均坐标被视为其重心，以红色点表示重心为参考)，根据从重心到关节点的平均距离将邻域划分为三个子集，分别为：根节点自身子集、距离重心节点较根节点近的子集、距离重心节点较根节点远的子集。

由于人运动时，关节点是以局部小组为单位移动的，本文中根据节点到根节点间的距离来划分邻域集。通过扩大整个邻域子集，进而关联根节点与更远节点之间的局部关系来加强身体各部分内部关节信息的关联，使模型能更敏感地感知身体局部信息，从而提升动态多尺度编码器提取特征的准确度。

1.2 NLNet

在深度神经网络中为了捕获长距离依赖，通常通过堆叠卷积层实现，这种方法虽然能够增大感受野，但计算效率低、建模困难、优化困难。而用于图像分类和语义分割的NLNet(non-local neural network)[22-23]不局限于相邻点，通过计算任意两个位置之间的交互直接捕捉远程依赖关系，摒弃了距离概念。NLNet的基本原理是先计算某点(点为向量，维度是通道数)与其余所有点的相似度，相似度越大，对最终的结果贡献就越大。对相似度进行归一化，得到各个点的权重。权重与对应点的特征映射值相乘，再与先前输入的点的特征相加就得到了包含全局信息的特征。

当人体在散步时，胳膊和腿上的关节点做出的贡献远大于其他关节点，于是文中利用非局部神经网络的优势，首次将非局部网络用于运动预测中计算关节点间的相互关系，使得网络在一个时空域中能更好地关注有贡献的关节点，并解决时空图卷积神经网络的接受域有限的问题。通过计算关节点之间的相似度得到时空中全部关节点在一项运动中对于某一关节点的作用，从而降低运动预测的误差。

1.3 GRU模型

GRU(gated recurrent unit)能够解决RNN(recurrent neural network)易出现的长期记忆依赖和反向传播中的梯度消失等问题，较LSTM(long short-term memory)有更少的输入和更简单的网络结构，能够大幅提高训练效率而不降低训练精度[24]。

GRU[25]更适合对长时动作的建模是因为它在RNN的内部设置了用于信息处理的两个门，分别是重置门(r_t)和更新门(z_t)。其中重置门(r_t)控制候选状态($\tilde{h}_t$)从上一时刻的状态(h_{t-1})中得到的信息度；更新门(z_t)使用$1-z_t$和z_t分别控制当前状态h_t从上一时刻状态(h_{t-1})中需要保留信息的力度和从候选状态($\tilde{h}_t$)中需要更新信息的度。计算过程如式(1)~式(4)所示：

$$r_t=\sigma(W_r[x_t,h_{t-1}]+b_r) \tag{1}$$

$$z_t=\sigma(W_z[x_t,h_{t-1}])+b_z \tag{2}$$

$$\tilde{h}_t=\tanh(W_h[x_t,x_t\odot h_{t-1}]+b_h) \tag{3}$$

$$h_t=(1-z_t)\odot\tilde{h}_t+z_t\odot h_{t-1} \tag{4}$$

其中，σ表示sigmoid激活函数；h_{t-1}表示上一时刻状态；x_t表示当前时刻的输入；$\tilde{h}_t$表示控制候选状态，以建立当前输入x_t和上一时刻状态h_{t-1}之间的联系；h_t表示隐藏状态。GRU模型的整体结构如图2所示。

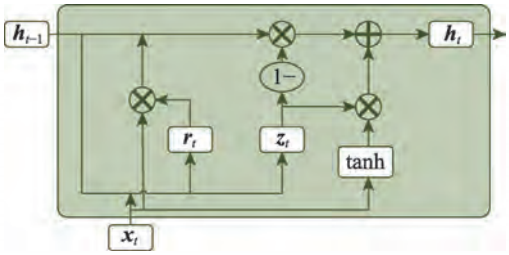

图2 GRU模型

Fig.2 Model of GRU

2 本文方法

2.1 编-解码网络

提出的人体运动预测多尺度全局自适应注意力图卷积算法(multiscale global adaptive attention graph neural network，MG-GNN)如图3所示，由编码网络和解码网络两部分组成。输入的骨架序列经过编码器准确地提取丰富的运动特征后送入解码器精确地预测骨架未来运动姿势。编码器具体结构如图4所示，由级联的多尺度全局自适应单元块(multiscale global adaptive unit，MGaU)组成。MGaU由多距离全局自适应注意力时空图卷积特征提取块(multi-distance global adaptive attention spatial temporal graph convolutional feature extraction block，MGST-FEB)和跨尺度融合块(cross-scale fusion block，CS-FB)组成。其中MGST-FEB采用了本文提出的多距离分区策略和全局自适应注意力时空卷积网络，能够提取更丰富

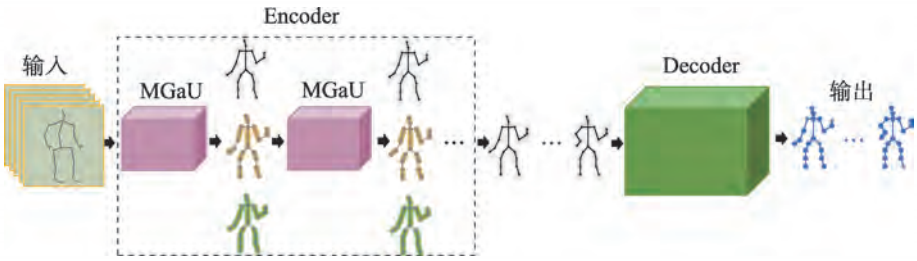

图3 MG-GNN网络的总体框架

Fig.3 General framework of MG-GNN

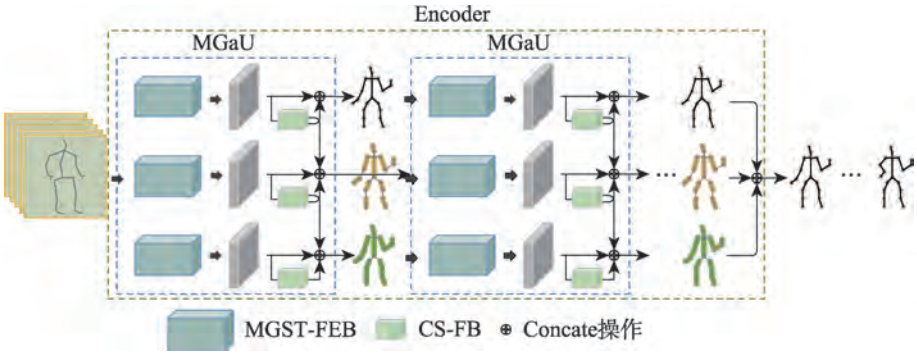

图4 编码器模型

Fig.4 Encoder model

的单一尺度的运动信息,CS-FB将粗细尺度信息进行融合,利用粗尺度对细尺度的指导以及细尺度对粗尺度的补充获得更加精细的运动特征。解码器具体结构如图5所示,采用基于多距离全局自适应注意力图卷积的门控循环单元(MGG-GRU),既利用可训练图增强状态传播,又利用门控循环单元结构使用残差连接增强预测结果。

2.2 多距离全局自适应注意力时空图卷积特征提取块(MGST-FEB)

编码网络由级联的MGaU块组成,而MGaU块中包含MGST-FEB和CS-FB两个模块,本文提出的MGST-FEB进行了两处改进:多距离分区策略和全局自适应注意力时空卷积图网络。解码网络由MGG-GRU组成,通过G-GRU采用基于多距离全局自适应注意力图卷积的门控循环单元改进得到MGG-GRU。

2.2.1 多距离分区策略(MD)

原始ST-GCN网络模型中的图卷积仅仅通过聚合相邻节点间的信息提取骨架关节点信息,且主要使用三种分区策略:单标签、距离分区和空间结构分区提取相邻节点之间的关系。但此三种分区策略仅

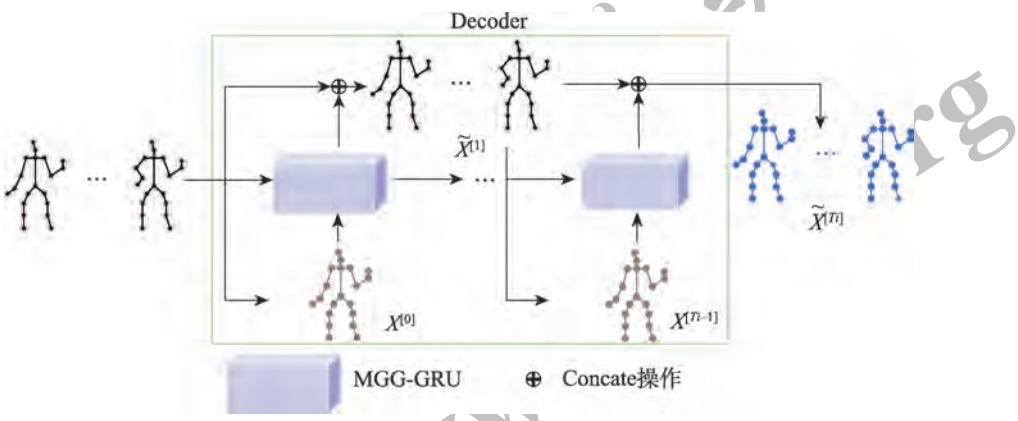

图5 解码器模型

Fig.5 Decoder model

仅考虑到了相邻关节点之间的关系，每个节点只将自己的信息传播给邻居节点，这会导致节点感受野较小，不利于获取长距离的连接信息。且人体活动时以局部小组为单位移动，故未能考虑到人身体部位内部关节点之间的联系对运动预测的重要性。

ST-GCN模型在二维卷积的基础上，通过重新定义的采样函数和权重函数即可构造空间上的图卷积公式为：

$$f_{\text{out}}(\boldsymbol{v}_{ti})=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{\boldsymbol{Z}_{ti}(\boldsymbol{v}_{tj})}f_{\text{in}}(\boldsymbol{v}_{tj})\cdot\boldsymbol{\omega}'(\boldsymbol{l}_{ti}(\boldsymbol{v}_{tj})) \tag{5}$$

其中，归一化项 $\boldsymbol{Z}_{ti}(\boldsymbol{v}_{tj})=|\{\boldsymbol{v}_{tk}|\boldsymbol{l}_{ti}(\boldsymbol{v}_{tk})=\boldsymbol{l}_{ti}(\boldsymbol{v}_{tj})\}|$ 为相应子集的基数，可平衡不同子集对输出的贡献；$\boldsymbol{v}_{tj}$ 是采样函数 P；$\boldsymbol{\omega}'(\boldsymbol{l}_{ti}(\boldsymbol{v}_{tj}))$ 是权重函数 $\boldsymbol{W}$ 。

图的时间信息是采用连续帧之间连接相同的关节点构建的，因此将空间的邻域概念扩展到包含时间连接的关节点的时间域。故采样函数可以定义为：

$$\boldsymbol{B}(\boldsymbol{v}_{ti})=\{\boldsymbol{v}_{qj}|\boldsymbol{d}(\boldsymbol{v}_{tj},\boldsymbol{v}_{ti})\leqslant\boldsymbol{K},|\boldsymbol{q}-\boldsymbol{t}|\leqslant\boldsymbol{\Gamma}/2\} \tag{6}$$

其中，T 为时间邻域的范围，即时间内核的大小。

因为时间轴的有序性，直接修改 $\boldsymbol{v}_{ti}$ 单帧关节点的映射 $\boldsymbol{l}_{ti}(\boldsymbol{v}_{tj})$，即可根据 $\boldsymbol{v}_{ti}$ 得出一个时空邻域。故权重函数定义为：

$$\boldsymbol{l}_{\text{ST}}(\boldsymbol{v}_{qj})=\boldsymbol{l}_{ti}(\boldsymbol{v}_{tj})+(\boldsymbol{q}-\boldsymbol{t}+\lfloor\boldsymbol{\Gamma}/2\rfloor)\boldsymbol{K} \tag{7}$$

由于观察人体运动例如“踢腿”“打电话”等动作是以人体部件为小组运动的，而ST-GCN算法中的三种策略的邻域均不能将人体部件关节点包含在里面，故为了充分发挥ST-GCN网络在时空域上对提取骨架关节点集成信息的重要性，本文提出了多距离分区策略。该策略的邻域能够涵盖人体运动的部件，并且能在单个帧中使用再扩展至空间-时间域，更加完整地提取运动特征，从而提高了每一个节点的感受野，更加差异化地学习不同节点的特征。故文中选取使用 $D=2$ 的相邻区域 $\boldsymbol{B}(\boldsymbol{v}_{ti})$ 设置采样函数 P，当 D 大于2时，节点邻域会超越身体部件，这导致网络提取特征时加入噪声，使得网络性能下降。多距离分区策略将根节点的邻域分成3个子集（如图6所示）：(1)距离根节点为0的子集（绿色）；(2)距离根节点子集为1的子集（蓝色）；(3)距离根节点为2的子集（黄色）。对每个子集的关节点赋予一种权重。则权重函数可以通过关节点之间的距离定义，每个子集的权重为：

$$\boldsymbol{L}_{ti}(\boldsymbol{v}_{ti})=\begin{cases}0, d=0\\1, d=1\\2, d=2\end{cases} \tag{8}$$

图6　多距离分区策略

Fig.6　Multi-distance partition strategy

其中，d 为节点到根节点的距离（跨越关节点的个数）。

采用类似于图卷积的表达形式[18]，单帧内关节的自连接由单位矩阵 $\boldsymbol{I}$ 和表示体内连接的相邻矩阵 $\boldsymbol{A}$ 表示，在单帧情况下可以使用以下公式实现：

$$f_{\text{out}}=\sum_{j}\boldsymbol{\Lambda}_{j}^{-\frac{1}{2}}\boldsymbol{A}_{J}\boldsymbol{\Lambda}_{J}^{-\frac{1}{2}}f_{\text{in}}\boldsymbol{W} \tag{9}$$

对于具有多个子集的分区策略，其中邻接矩阵被拆分成几个矩阵 $\boldsymbol{A}_j$，即 $\boldsymbol{A}+\boldsymbol{I}=\sum_j\boldsymbol{A}_j$。在该策略分区中，$\boldsymbol{A}_0=\boldsymbol{I}$ 和 $\boldsymbol{A}_1+\boldsymbol{A}_2+\boldsymbol{A}_3=\boldsymbol{A}$ 。

多距离分区策略考虑到人体运动是以局部小组为单位移动，故采样函数 P 使用 $D=2$ 的邻域 $\boldsymbol{B}(\boldsymbol{v}_{ti})$，扩大了整个邻域集，通过提取根节点与更远关节点之间的信息，加强身体各部分内部关节点之间的联系，从而提高模型对身体局部的敏感性，进一步降低预测误差。并且本文将多距离分区策略与原ST-GCN中的三种分区策略进行了对比，实验结果在3.3.1小节中，通过实验验证的方法再次证明本文提出的多距离分区策略的优越性。

2.2.2　全局自适应注意力时空图卷积网络（GaST-GCN）

由于ST-GCN中卷积核感受野的限制，导致模型不能捕获某一关节点和全部关节点之间的时空信息，且不能区分对某一运动有突出贡献的其他关节点，不利于运动特征的提取。由于NLNet[22,26]的全局自适应注意力可以通过计算任意两个关节点之间的交互直接捕捉远程依赖，不局限于邻域，可捕获更多关节点的时空全局信息，并且网络通过相似度的大小给予不同关节点不同的关注度，减少不相关关节点对于网络的干扰。与NLNet网络不同之处在于（如图7所示）本文提出的GaST-GCN在NLNet网络中采用时空卷积块对网络得到的相关关系进行操作，并且将残差运用到时空卷积操作之后，这样使得

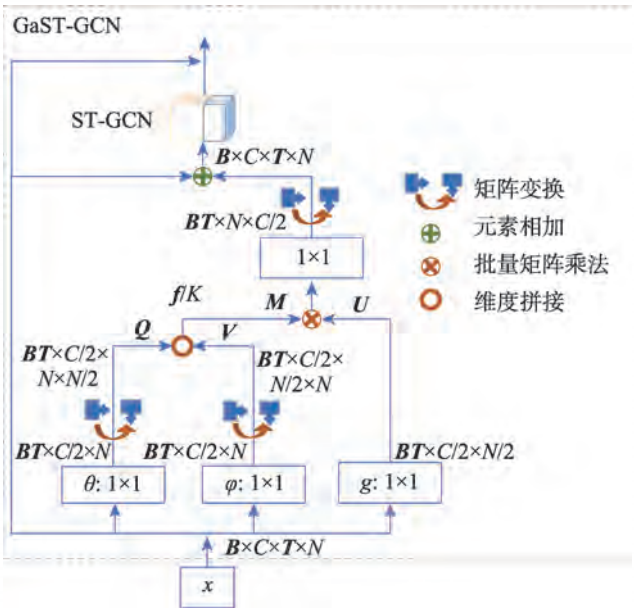

图7 GaST-GCN模型

Fig.7 Model of GaST-GCN

编码网络能够提取到时空卷积网络邻域之外的信息，同时不用再增加额外的计算量。自适应是通过每一批次中关节点之间的相对位置的不同而体现。

GaST-GCN中的非局部关节点注意力可定义为：

$$\boldsymbol{x}_i^{(l+1)}=\boldsymbol{x}_i^{(l)}+\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{f}(\boldsymbol{x}_i^{(l)},\boldsymbol{x}_j^{(l)})\cdot\boldsymbol{g}(\boldsymbol{x}_i^{(l)}) \tag{10}$$

其中，$\boldsymbol{x}_i^{(l)}$ 和 $\boldsymbol{x}_i^{(l+1)}$ 分别为输入特征和输出特征；K 为归一化参数（人体关节点总数）；函数 $f(x_i,x_j)=\mathrm{ReLU}(\boldsymbol{w}_f^{\mathrm{T}}[\theta(x_i),\phi(x_j)])$ 用于学习节点 i 和其他节点 j 之间的密切程度；函数 g 用于计算骨架第 j 个关节点的位置特征。由式(10)可知，GaST-GCN中的操作考虑了当前关节点与特征空间中所有关节点的联系，因此，可有效地捕捉到骨架的长时依赖关系。

2.3 基于多距离全局自适应注意力图卷积的门控循环单元(MGG-GRU)

结合全局自适应注意力的优势并在图卷积中使用多距离分区策略，设一个基于多距离全局自适应注意力图卷积的门控循环单元模型(MGG-GRU)。其结构如图8所示，GRU的隐藏状态是在MGG-GRU指导下学习和更新的。MGG-GRU有两个输入状态，分别是上一时刻的状态 $\boldsymbol{h}_{t-1}\in\mathbf{R}^{M\times d}$（使用 M 个关节点和 d=3表示时间 t 的3D姿势）和基于三位骨架的信息状态 $\boldsymbol{SI}_t\in\mathbf{R}^{M\times d}$，计算过程如式(11)~式(14)。

$$\boldsymbol{r}_t=\sigma(\boldsymbol{r}_{\mathrm{in}}(\boldsymbol{SI}_t)+\boldsymbol{r}_{\mathrm{hid}}(\boldsymbol{A}_H\boldsymbol{h}_{t-1}\boldsymbol{W}_H)) \tag{11}$$

$$\boldsymbol{z}_t=\sigma(\boldsymbol{z}_{\mathrm{in}}(\boldsymbol{SI}_t)+\boldsymbol{z}_{\mathrm{hid}}(\boldsymbol{A}_H\boldsymbol{h}_{t-1}\boldsymbol{W}_H)) \tag{12}$$

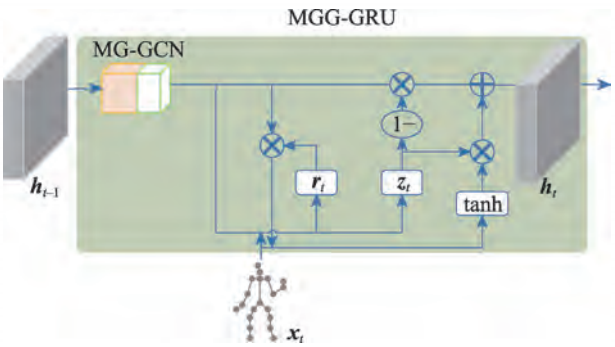

图8 MGG-GRU模型

Fig.8 Model of MGG-GRU

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{h}_{\mathrm{in}}(\boldsymbol{SI}_t)+\boldsymbol{r}_t\odot\boldsymbol{h}_{\mathrm{hid}}(\boldsymbol{A}_H\boldsymbol{h}_{t-1}\boldsymbol{W}_H)) \tag{13}$$

$$\boldsymbol{h}_t=\boldsymbol{z}_t\odot\boldsymbol{h}_{t-1}+(1-\boldsymbol{z}_t)\odot\tilde{\boldsymbol{h}}_t \tag{14}$$

其中，$\boldsymbol{A}_H\in\mathbf{R}^{M\times d}$ 用骨架图初始化，是内置图的邻接矩阵；$\boldsymbol{r}_{\mathrm{in}}(\cdot)$、$\boldsymbol{r}_{\mathrm{hid}}(\cdot)$、$\boldsymbol{z}_{\mathrm{in}}(\cdot)$、$\boldsymbol{z}_{\mathrm{hid}}(\cdot)$、$\boldsymbol{h}_{\mathrm{in}}(\cdot)$、$\boldsymbol{h}_{\mathrm{hid}}(\cdot)$ 是可训练线性映射；$\boldsymbol{W}_H$ 是可训练权重。

3 实验结果与讨论

采用PyTorch-1.8.0深度学习框架，编程语言为Python_3.8，在Windows操作系统下实现，GPU为ATX-4000，CUDA版本为11.2。

3.1 实验数据集和评价指标

Human 3.6M(H3.6M)数据集包含7名受试者，每个受试者执行15种不同的动作。每个人有32个关节点，将关节点位置通过指数图转换为坐标图，只使用非零关节点，并沿时间轴对所有序列进行两次下采样。选取S1、S6、S7、S8、S9、S11共6名受试者为训练集，选取S5为测试集。

CMU Mocap(CMU)数据集是由情景&场景、人类互动、运动、与环境的互动、体育活动&运动5个主要动作类别构成的常用的人体姿势预测数据集。通过非零指数图给每个受试者保留26个关节点。由于所提算法没有使用“人机交互”类别的数据，并且需要为每个行为提供足够的数据来训练模型，排除了包含少于6个试验的行为类别。故从“通信手势和信号”类别中选择交通方向和篮球信号，从“运动”类别中选择跑步、步行和跳跃，从“常见行为和表达”类别中选择清洗窗户，从“体育活动&运动”类别中选择篮球和足球，共8个动作类别。对数据集的处理方式和评估方法都和H3.6M相同。

3DPW数据集[27]由超过510 000帧的三维姿势组成的大型数据集，用于挑战性的室内和室外活动。本文采用官方设置的训练集和测试集。

评价指标采用与基线算法相同的角空间中产生的运动和地面真实值之间的平均角误差（mean angular error，MAE）定量评估各种方法之间的性能。通过本文算法与当前流行的算法进行比较，展示本文算法的性能，同时通过组合各个模块比较，以展示所提算法的有效性。

Kinetics-skeleton数据集是建立在大规模动作识别数据集Kinetics（Kinetics人类动作数据集）上的。Kinetics是最大的无约束动作识别数据集，包含约30万个从YouTube检索的视频片段。这些视频涵盖了多达400个人类动作，从日常活动和体育场景到复杂的互动动作，视频中的每个片段都持续了大约10 s。Yan等人通过在Kinetics数据集上使用OpenPose获得每帧上18个骨骼点的二维坐标（X,Y）和置信度分数C，并保留每帧内平均置信度最高的2个人的关节，并为每个动作选择300帧作为动作骨骼序列，从而建立了Kinetics骨骼数据集。该数据集提供了一个由24万个片段组成的训练集和一个由2万个片段组成的验证集。为了便于比较，本文在训练集上训练了模型，并在验证集上验证了模型的性能。文中使用Top-1和Top-5的准确度指标进行验证。

3.2 实验结果分析

为了验证MG-GNN模型的性能，在H3.6M数据集、3DPW数据集和CMU数据集上进行模型的训练和测试，采用与早期研究[6]相同的子序列进行评估，并采用与DMGNN算法相同的优化器和损失函数。总迭代次数为4万次，学习率为0.000 05，批次大小设置为32。

3.2.1 在H3.6M数据集上的实验结果

短期运动预测（预测500 ms以内的未来姿势）：将本文算法MG-GNN与其他算法在H3.6M数据集上预测400 ms运动的结果进行比较。结果如表1和表2所示。首先可以看出在H3.6M数据集上的所有动作的短期运动预测误差均值在160 ms、320 ms、400 ms均好于其他算法，且在H3.6M数据集的4类代表性动作“散步”“进食”“吸烟”“讨论”上的运动误差均值如图9所示，这4类代表性动作的预测结果可视化如图10所示；其次从表1中可以看出：MG-GNN与基线DMGNN相比，由于使用的多距离分区策略丰富了网络对于身体各部分关节点之间的信息的提取，使得MG-GNN除了与基线相同的在局部周期性行为“进食”“吸烟”方面的预测良好之外，在“散步”上也获得了很低的MAE；最后由于加入全局自适应注意力，有效提取了对某一项运动有贡献的时空中的其他关节点信息，使得MG-GNN在非周期性的行为“讨论”上取得了较原算法更精确的预测。

但是结果与其他算法相比仍不够精确，由于此模型对于非周期性运动的预测需要在更长时间序列上得出较为精确的预测，表3证实了这一想法。表2给出了MG-GNN与一些具有代表性的方法和最新的方法在H3.6M数据集中剩余的11类动作的比较，结果表明MG-GNN在大多数动作类中能够达到较精确的预测，而在“讨论”“购买”“拍照”短期动作上表现不尽如人意。原因在于这些非周期动作，在时间轴上的变化大，基线网络和本文新提出的网络中都没有在时序方面的增强机制，导致在这些动作上表现较差。

长期运动预测（超过500 ms的未来姿势）：由于动作变化和非线性，长期运动预测具有很大的挑战性。表3显示了本文算法与其他算法在H3.6M数据集中具有代表性的4类动作上560 ms和1 000 ms的预测结果。从结果首先可以看出长期运动预测在560 ms的误差均值比其他方法都好，在1 000 ms的误差均值也好于原始算法，但是在短期运动预测的好的结果并没有在长期运动预测中延续，MG-GNN和DMGNN在1 000 ms时的“散步”表现均不理想，这是由于在跨尺度融合块中并未使用有利于长时依赖的

表1 不同方法在H3.6M数据集上短期运动预测的MAE比较

Table 1 MAE comparison of short-term motion prediction on H3.6M dataset by different methods

Method	Walking				Eating				Smoking				Discussion			
	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms
Res-sup.[28]	0.27	0.46	0.67	0.75	0.23	0.37	0.59	0.73	0.32	0.59	1.01	1.10	0.30	0.67	0.98	1.06
CSM[29]	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[9]	0.18	0.32	0.49	0.56	0.17	0.31	0.52	0.62	0.22	0.41	0.84	0.79	**0.20**	**0.51**	0.79	0.86
DMGNN[15]	0.18	0.31	0.49	0.58	0.17	0.30	0.49	0.59	**0.21**	**0.39**	0.81	0.77	0.26	0.65	0.92	0.99
Sybio-GNN[30]	**0.17**	0.31	0.50	0.60	0.16	0.29	0.48	0.60	**0.21**	0.40	**0.76**	0.80	0.21	0.55	**0.77**	**0.85**
MG-GNN	0.18	**0.29**	**0.46**	**0.54**	**0.15**	**0.27**	**0.45**	**0.57**	**0.21**	**0.39**	0.80	**0.74**	0.21	0.58	0.84	0.91

表2 不同方法在H3.6M数据集的其他11个动作类上的短期运动预测的MAE比较

Table 2 MAE comparison of short-term motion prediction on other 11 action classes of H3.6M dataset by different methods

Method	Directions				Greeting				Phoning				Posing			
	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms
Res-sup.[28]	0.41	0.64	0.80	0.92	0.57	0.83	1.45	1.60	0.59	1.06	1.45	1.60	0.45	0.85	1.34	1.56
CSM[29]	0.39	0.60	0.80	0.91	0.51	0.82	1.21	1.38	0.59	1.13	1.51	1.65	0.29	0.60	1.12	1.37
Traj-GCN[9]	0.26	0.45	0.70	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[15]	0.25	0.44	0.65	0.71	0.36	0.61	0.94	**1.12**	0.52	0.97	1.29	1.43	0.25	0.44	**0.65**	**0.71**
Sybio-GNN[30]	0.23	**0.42**	**0.57**	0.65	**0.35**	**0.60**	0.95	1.15	**0.48**	**0.80**	1.28	1.41	0.18	0.45	0.97	1.20
MG-GNN	**0.22**	0.45	**0.57**	**0.64**	**0.35**	**0.60**	**0.93**	1.13	0.51	0.90	**1.25**	**1.37**	**0.17**	**0.39**	0.79	0.97

Method	Purchases				Sitting				Sitting Down				Taking Photo			
	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms
Res-sup.[28]	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	0.90	1.02
CSM[29]	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[9]	0.42	0.66	1.04	1.12	0.29	0.45	0.82	0.97	0.30	0.63	0.89	1.01	0.15	0.36	0.59	0.72
DMGNN[15]	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
Sybio-GNN[30]	**0.40**	**0.60**	0.97	1.04	**0.24**	0.41	0.77	0.95	**0.28**	0.60	0.89	0.99	**0.14**	**0.32**	**0.53**	**0.64**
MG-GNN	**0.40**	0.61	**0.95**	**1.02**	**0.24**	**0.40**	**0.75**	**0.93**	0.29	**0.59**	**0.85**	**0.96**	0.15	0.34	0.55	0.66

Method	Waiting				Walking Dog				Walking Together				Average			
	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms	80 ms	160 ms	320 ms	400 ms
Res-sup.[28]	0.32	0.63	1.07	1.26	0.52	0.89	1.25	1.40	0.27	0.53	0.74	0.79	0.40	0.69	1.04	1.18
CSM[29]	0.30	0.62	1.09	1.30	0.59	1.00	1.32	1.44	0.27	0.52	0.71	0.74	0.38	0.68	1.01	1.13
Traj-GCN[9]	0.23	0.50	0.92	1.15	0.46	0.80	1.12	1.30	0.15	0.35	0.52	0.57	0.27	0.53	0.85	0.96
DMGNN[15]	0.22	0.49	1.10	1.10	0.42	0.72	1.16	1.34	0.15	0.33	0.50	0.57	0.27	0.52	0.83	0.95
Sybio-GNN[30]	0.22	0.48	0.87	1.06	0.42	0.73	1.08	1.22	0.16	0.33	0.50	0.56	**0.26**	0.49	0.79	0.92
MG-GNN	**0.21**	**0.47**	**0.82**	**0.97**	**0.40**	**0.68**	**1.05**	**1.20**	**0.14**	**0.30**	**0.46**	**0.50**	0.28	**0.48**	**0.77**	**0.87**

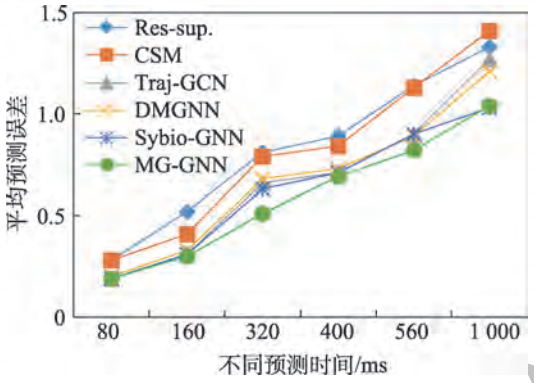

图9 H3.6M数据集上不同模型的平均角度误差

Fig.9 Mean angular error of different models on H3.6M dataset

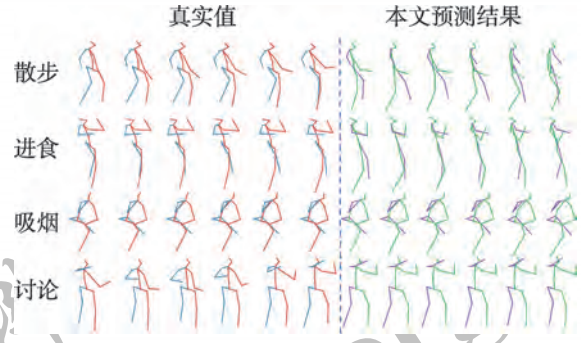

图10 H3.6M数据集的4类代表性动作的短期预测可视化结果

Fig.10 Short-term prediction qualitative results of 4 representative actions on H3.6M dataset

时间注意力机制而导致的。其次在“讨论”上的长期运动预测表现比短期运动预测的结果好,这一结果说明对于非周期性运动的预测在较长时间中会获得更好的结果。

3.2.2 在CMU数据集上的实验结果

在CMU数据集中挑选的8类动作上进行了短期和长期运动预测,并与其他算法进行了比较,结果如表4所示。与不同运动预测模型在不同预测时间的

表3 不同方法在H3.6M数据集上长期运动预测的MAE比较

Table 3 MAE comparison of long-term motion prediction of H3.6M dataset by different methods

Method	Walking		Eating		Smoking		Discussion		Average	
	560 ms	1 000 ms	560 ms	1 000 ms	560 ms	1 000 ms	560 ms	1 000 ms	560 ms	1 000 ms
Res-sup.[28]	0.93	1.03	0.95	1.08	1.25	1.50	1.43	1.69	1.14	1.33
CSM[29]	0.98	0.92	1.01	1.24	0.97	1.62	1.56	1.86	1.13	1.41
Traj-GCN[9]	0.65	0.67	0.76	1.12	0.87	1.57	1.33	1.70	0.90	1.27
DMGNN[15]	0.66	0.75	0.74	1.14	0.83	1.52	1.33	1.45	0.89	1.21
Sybio-GNN[30]	0.75	0.78	0.77	0.88	0.92	1.18	1.17	1.28	0.90	1.03
MG-GNN	0.65	0.77	0.78	1.15	0.71	1.21	1.14	1.05	0.82	1.04

表4 不同方法在CMU数据集上的短期和长期运动预测的MAE比较

Table 4 MAE comparison of short-term and long-term motion prediction of CMU dataset by different methods

Method	Basketball					Basketball Signal				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
Res-sup.[28]	0.49	0.77	1.26	1.45	1.77	0.42	0.76	1.33	1.54	2.17
CSM[29]	0.36	0.62	1.07	1.17	1.95	0.33	0.62	1.05	1.23	1.98
Traj-GCN[9]	0.33	0.52	0.89	1.06	1.71	0.11	0.20	0.41	0.53	1.00
DMGNN[15]	0.30	**0.46**	0.89	1.11	1.66	0.10	0.17	0.31	0.41	1.26
Sybio-GNN[30]	0.32	0.48	0.91	1.06	**1.47**	0.12	0.21	0.38	0.49	0.94
MG-GNN	**0.28**	**0.46**	**0.86**	**1.03**	1.50	**0.09**	**0.16**	**0.30**	**0.40**	**0.91**
Method	Directing Traffic					Jumping				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
Res-sup.[28]	0.31	0.58	0.94	1.10	2.06	0.57	0.86	1.76	2.03	2.42
CSM[29]	0.26	0.58	0.91	1.04	2.08	0.38	0.60	1.36	1.58	2.05
Traj-GCN[9]	**0.15**	0.32	**0.52**	**0.60**	2.00	**0.31**	**0.49**	**1.23**	**1.39**	1.80
DMGNN[15]	**0.15**	**0.30**	0.57	0.72	1.98	0.37	0.65	1.49	1.71	1.79
Sybio-GNN[30]	0.20	0.41	0.75	0.87	**1.84**	0.32	0.55	1.40	1.60	1.82
MG-GNN	**0.15**	0.32	0.55	0.66	1.91	0.35	0.61	1.45	1.65	**1.72**
Method	Running					Soccer				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
Res-sup.[28]	0.32	0.48	0.65	0.74	1.00	0.29	0.50	0.87	0.98	1.73
CSM[29]	0.28	0.43	0.54	0.57	0.69	0.28	0.48	0.79	0.90	1.58
Traj-GCN[9]	0.33	0.55	0.73	0.74	0.95	**0.18**	**0.29**	**0.61**	**0.71**	1.40
DMGNN[15]	0.19	0.31	0.47	0.49	0.64	0.22	0.32	0.79	0.91	1.54
Sybio-GNN[30]	0.21	0.33	0.53	0.56	0.65	0.19	0.32	0.66	0.78	**1.32**
MG-GNN	**0.18**	**0.29**	**0.39**	**0.41**	**0.56**	0.19	**0.29**	0.70	0.82	1.38
Method	Walking					Washing Window				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
Res-sup.[28]	0.35	0.45	0.59	0.64	0.88	0.31	0.47	0.74	0.93	1.37
CSM[29]	0.35	0.44	0.46	0.51	0.77	0.30	0.47	0.79	1.00	1.39
Traj-GCN[9]	0.33	0.45	0.49	0.53	0.61	0.22	0.33	0.57	0.75	1.20
DMGNN[15]	0.30	0.34	0.38	0.43	0.60	0.20	0.27	0.62	0.81	1.09
Sybio-GNN[30]	**0.26**	**0.32**	**0.35**	0.39	**0.52**	0.22	0.33	0.55	0.73	1.05
MG-GNN	0.29	0.33	**0.35**	**0.38**	**0.52**	**0.19**	**0.26**	**0.51**	**0.66**	**1.02**

所有运动的平均预测误差趋势如图11所示。从结果首先可以看到,MG-GNN的长短期运动预测误差均值在160 ms、320 ms、400 ms、1 000 ms均好于其他算法;其次可以看出MG-GNN在除了“跳跃”以外的动作类上都取得了良好的结果;最后可以看出在“跳跃”上MG-GNN算法与原始算法相比预测误差都有所下降,但是两者均差于Traj-GCN[9]方法。是由于人体在跳跃的时候,在时空中的变化幅度较大,关节点

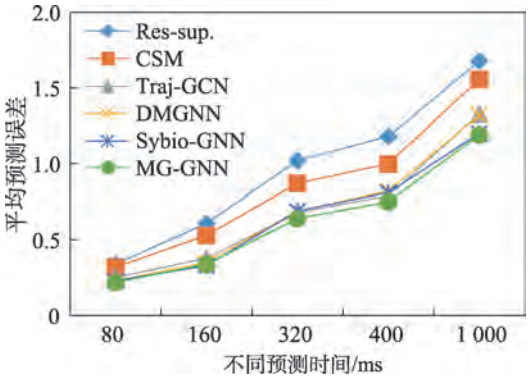

图11 CMU数据集上的不同模型的平均预测误差

Fig.11 Average prediction errors of different models on CMU dataset

重叠率较高,此时对于模型提取关节点的长时动作依赖要求更高。MG-GNN算法只在提取运动特征的模块中嵌入了全局自适应注意力,较难胜任在时空中变化幅度较大的动作预测。

3.2.3 在3DPW数据集上的实验结果

为了验证本文方法,在3DPW数据集上对于短期和长期运动预测的平均MAE如表5所示,本文所提模型在短期和长期运动预测的平均MAE均好于其他模型再次证明模型的鲁棒性。

表5 不同方法在3DPW数据集上的短期和长期运动预测的平均MAE比较

Table 5 Average MAE comparison of short-term and long-term motion prediction of 3DPW dataset by different methods

Method	Average MAE				
	200 ms	400 ms	600 ms	800 ms	1 000 ms
Res-sup.[28]	1.95	2.37	2.46	2.51	2.53
CSM[29]	1.24	1.85	2.13	2.23	2.26
Traj-GCN[9]	0.64	0.94	**1.11**	1.22	1.27
DMGNN[15]	0.62	0.93	1.14	1.23	1.26
MG-GNN	**0.61**	**0.90**	**1.11**	**1.21**	**1.24**

3.3 消融实验

3.3.1 在Kinetics-skeleton数据集上的消融实验

由于本文中的MD策略和GaST-GCN模块是对动作识别网络ST-GCN的改进部分,为了验证改进的效果,本文在Kinetics-skeleton这个动作识别的常用数据集上进行了实验。多距离分区策略模型与其他模型在Kinetics-skeleton数据集上的性能比较见表6。从表中的数据来看,与ST-GCN[18]相比,MD的Top-1和Top-5分别提高了1.47个百分点和2.01个百分点。与原算法中使用的距离分区策略相比,Top-1和Top-5分别提高了3.07个百分点和3.51个百分点。

表6 不同分区策略与ST-GCN模型的准确性比较

Table 6 Accuracy comparison of different partition strategies with ST-GCN model 单位:%

Partition strategy	Top-1	Top-5
Uni-labeling	19.30	37.40
Distance Partitioning	29.10	51.30
Spatial Configuration	29.90	52.20
Spatial Configuration*[18]	30.70	52.80
Multi-distance Partition (MD)	**32.17**	**54.81**

在动作识别数据集Kinetics-skeleton上将全局自适应注意力运用到ST-GCN模型的时空块的结果如图12所示。当采用1到2个全局自适应注意力层时,Top-1和Top-5上升;当采用3到9个全局自适应注意力层时,Top-1和Top-5下降。故采用0或1个全局自适应注意力时融合不充分,而更多的全局自适应注意力层往往会融合多余的信息,使模型混乱。因此,在运动预测模型中,本文也使用了两层的全局自适应注意力。

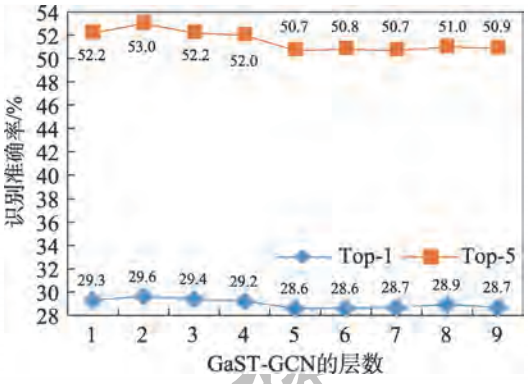

图12 不同GaST-GCN层数在Kinetics-skeleton数据集上的精度比较

Fig.12 Accuracy comparison of different layers of GaST-GCN on Kinetics-skeleton dataset

3.3.2 在H3.6M数据集上的消融实验

为了验证本文设计的多距离分区策略和全局自适应注意力在DMGNN算法上的有效性,在H3.6M数据集的4类代表性动作上进行短期预测和长期预测。

实验结果如表7所示,在仅使用多距离分区策略后,DMGNN提取人体关节点特征时考虑身体部位中的关节点之间的联系,从结果来看,在H3.6M数据集的4类代表性动作上的长短期预测结果略好于原算法;在仅使用全局自适应注意力后,DMGNN在提取

表7 加入不同模块网络的效果

Table 7 Effect of network when adding different modules

Method	Walking					Eating				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
DMGNN[15]+MD	0.179	0.309	0.485	0.576	0.799	0.169	0.295	0.472	0.587	**1.117**
DMGNN[15]+GaST-GCN	0.178	0.303	0.478	0.557	**0.743**	0.154	0.299	0.483	0.585	1.182
DMGNN[15]+MGG-GRU	0.181	0.302	0.500	0.561	0.783	0.158	0.281	0.471	0.588	1.166
DMGNN[15]+MGST-GCN	0.176	0.296	0.470	0.540	0.769	0.159	0.289	0.469	0.577	1.157
DMGNN[15]+MGST-GCN+MGG-GRU(MG-GNN)	**0.173**	**0.293**	**0.459**	**0.538**	0.773	**0.152**	**0.272**	**0.451**	**0.565**	1.150
Method	Smoking					Discussion				
	80 ms	160 ms	320 ms	400 ms	1 000 ms	80 ms	160 ms	320 ms	400 ms	1 000 ms
DMGNN[15]+MD	0.215	0.396	0.852	0.777	1.392	0.246	0.673	0.898	0.988	1.176
DMGNN[15]+GaST-GCN	0.212	0.391	0.833	0.755	1.193	0.252	0.619	0.896	0.987	1.107
DMGNN[15]+MGG-GRU	0.217	0.406	0.807	0.774	1.296	0.248	0.619	0.881	0.942	1.085
DMGNN[15]+MGST-GCN	0.213	0.392	0.808	0.743	1.265	0.235	0.599	0.857	0.978	1.162
DMGNN[15]+MGST-GCN+MGG-GRU(MG-GNN)	**0.211**	**0.391**	**0.799**	**0.741**	**1.054**	**0.208**	**0.584**	**0.843**	**0.912**	**1.046**

人体关节点时考虑了全部有关的时空中的关节点信息，在H3.6M数据集的4类代表性动作上的长短期预测结果略好于原算法；仅使用改进的解码网络后，由于加强了动作状态的传播性能，改进后的解码器对于运动预测的结果略好于原算法；从表7中的模块组合实验结果可以看出当3个模块同时添加时，增强了身体部位内部关节点间的信息以及与动作有关的全部时空中的关节点某信息，证实了所提算法的有效性。

4 结束语

为实现准确、鲁棒的人体运动预测，提出了一种多尺度全局自适应图神经网络的人体运动预测算法。设计了一种多距离分区策略，可帮助网络更好地提取身体关节点各部分关节点之间的联系信息，以增强特征中的运动信息。在ST-GCN的时空块中嵌入全局自适应注意力，提取对某一运动贡献度高的全局关节点信息，突破ST-GCN算法感受野的局限性。实验表明，所提的网络模型优于当前运动预测性能较好的算法模型。本文算法的预测速度和预测精度在短期预测“讨论”“跳跃”和长期预测“散步”上仍有待提高，拟在多尺度融合块中加入时间自适应模块，高效灵活地捕捉非周期性运动和长期预测的时间关联性；拟在解码网络的GRU中加入软注意力机制，通过选择性地忽略部分信息来对其余信息进行重加权聚合计算，提高网络对非周期性运动和长期运动的预测能力。同时收集更多运动预测的数据集，以便进行更全面的训练和预测，进一步提高算法的预测效果。

参考文献：

[1] PADEN B, ČAP M, YONG S Z, et al. A survey of motion planning and control techniques for self-driving urban vehicles[J]. IEEE Transactions on Intelligent Vehicles, 2016, 1 (1): 33-55.

[2] SANG H F, CHEN Z Z, HE D K. Human motion prediction based on attention mechanism[J]. Multimediatools and Applications, 2020, 79(9): 5529-5544.

[3] YANG H, YUAN C, ZHANG L, et al. STA-CNN: convolutional spatial-temporal attention learning for action recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 5783-5793.

[4] 邓辉, 徐杨. 融入注意力和密集连接的轻量型人体姿态估计[J]. 计算机工程与应用, 2022, 58(16): 265-273.
DENG H, XU Y. Lightweight human pose estimation based on attention and dense connection[J]. Computer Engineering and Applications, 2022, 58(16): 265-273.

[5] GUI L Y, WANG Y X, LIANG X, et al. Adversarial geometry-aware human motion prediction[C]//LNCS 11208: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 823-842.

[6] KUNDU J N, GOR M, BABU R V. BiHMP-GAN: bidirectional 3D human motion prediction GAN[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances

in Artificial Intelligence, Honolulu, Jan 27- Feb 1, 2019. Menlo Park: AAAI, 2019: 8553-8560.

[7] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems 29, Barcelona, Dec 5-10, 2016: 3844-3852.

[8] GUO X, CHOI J. Human motion prediction via learning local structure representations and temporal dependencies [C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, Jan 27-Feb 1, 2019. Menlo Park: AAAI, 2019: 2580-2587.

[9] MAO W, LIU M, SALZMANN M, et al. Learning trajectory dependencies for human motion prediction[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 9489-9497.

[10] MAO W, LIU M, SALZMANN M. History repeats itself: human motion prediction via motion attention[C]//LNCS 12359: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 474-489.

[11] ZHOU H, GUO C, ZHANG H, et al. Learning multiscale correlations for human motion prediction[C]//Proceedings of the 2021 IEEE International Conference on Development and Learning, Beijing, Aug 23-26, 2021. Piscataway: IEEE, 2021: 1-7.

[12] MAO W, LIU M, SALZMANN M, et al. Multi-level motion attention for human motion prediction[J]. International Journal of Computer Vision, 2021, 129(9): 2513-2535.

[13] MAO W, LIU M, SALZMANN M. Generating smooth pose sequences for diverse human motion prediction[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 13289-13298.

[14] WANG H, HO E S L, SHUM H P H, et al. Spatio-temporal manifold learning for human motions via long-horizon modeling[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 27 (1): 216-227.

[15] LI M, CHEN S, ZHAO Y, et al. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 214-223.

[16] DANG L, NIE Y, LONG C, et al. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 11447-11456.

[17] LI M, CHEN S, ZHAO Y, et al. Multiscale spatio-temporal graph neural networks for 3D skeleton-based motion prediction[J]. IEEE Transactions on Image Processing, 2021, 30: 7760-7775.

[18] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 7444-7452.

[19] 严春满, 王铖. 卷积神经网络模型发展及应用[J]. 计算机科学与探索, 2021, 15(1): 27-46.
YAN C M, WANG C. Development and application of convolutional neural network model[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(1): 27-46.

[20] 邓淼磊, 高振东, 李磊, 等. 基于深度学习的人体行为识别综述[J]. 计算机工程与应用, 2022, 58(13): 14-26.
DENG M L, GAO Z D, LI L, et al. Overview of human behavior recognition based on deep learning[J]. Computer Engineering and Applications, 2022, 58(13): 14-26.

[21] 何坚, 郭泽龙, 刘乐园, 等. 基于滑动窗口和卷积神经网络的可穿戴人体活动识别技术[J]. 电子与信息学报, 2022, 44(1): 168-177.
HE J, GUO Z L, LI L Y, et al. Human activity recognition technology based on sliding windowand convolutional neural network[J]. Journal of Electronics & Information Technology, 2022, 44(1): 168-177.

[22] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 7794-7803.

[23] 梁延禹, 李金宝. 多尺度非局部注意力网络的小目标检测算法[J]. 计算机科学与探索, 2020, 14(10): 1744-1753.
LIANG Y Y, LI J B. Small objects detection method based on multi-scale non-local attention network[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14 (10): 1744-1753.

[24] 盖杉, 王俊生. 基于深度学习的非局部注意力增强网络图像去雨算法研究[J]. 电子学报, 2020, 48(10): 1899-1908.

GAI S, WANG J S. Image raindrop algorithm research using nonlocal attention enhanced network based on deep learning [J]. Journal of Electronics & Information Technology, 2020, 48(10): 1899-1908.

[25] WANG B, ADELI E, CHIU H, et al. Imitation learning for human pose prediction[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 7124-7133.

[26] BUADES A, COLL B, MOREL J M. A non-local algorithm for image denoising[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, Jun 20-26, 2005. Washington: IEEE Computer Society, 2005: 60-65.

[27] VON MARCARD T, HENSCHEL R, BLACK M J, et al. Recovering accurate 3D human pose in the wild using IMUs and a moving camera[C]//LNCS 11214: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 614-631.

[28] MARTINEZ J, BLACK M J, ROMERO J. On human motion prediction using recurrent neural networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2891-2900.

[29] LI C, ZHANG Z, LEE W S, et al. Convolutional sequence to sequence model for human dynamics[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 5226-5234.

[30] LI M, CHEN S, CHEN X, et al. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3316-3333.

苟茹茹(1993—),女,甘肃定西人,硕士研究生,CCF学生会员,主要研究方向为人体动作识别、运动预测。

GOU Ruru, born in 1993, M.S. candidate, student member of CCF. Her research interests include action recognition and motion prediction.

杨文柱(1969—),男,河北顺平人,工学博士,教授,硕士生导师,CCF会员,主要研究方向为机器视觉、智能系统。

YANG Wenzhu, born in 1969, Ph.D., professor, M.S. supervisor, member of CCF. His research interests include machine vision and intelligent systems.

罗梓菲(1997—),女,四川成都人,硕士研究生,CCF学生会员,主要研究方向为语义分割。

LUO Zifei, born in 1997, M.S. candidate, student member of CCF. Her research interest is semantic segmentation.

原云峰(1995—),男,山西阳城人,硕士研究生,CCF学生会员,主要研究方向为动作识别。

YUAN Yunfeng, born in 1995, M.S. candidate, student member of CCF. His research interest is action recognition.