面向骨架手势识别的全局时空可变形网络

石东子[1]，林宏辉[1]，刘一江[1]，张 鑫[1, 2*]

（1. 华南理工大学 电子与信息学院，广州 510640；2. 人工智能与数字经济广东省实验室，广州 510640）

摘要 基于骨架序列进行手势识别关键在于如何融合时空信息提取可分辨性强的特征。该文提出关键点聚焦模块，通过全局上下文建模和不受限于固定形式的卷积方式，网络可以跨越多帧和不相关的关键点，在全局范围内自适应地聚合与手势动作密切相关的关键点信息，提取手势的时空特征。实验表明该方法在 ChaLearn2013 和 SHREC 数据集上得到的准确率可以达到 94.88% 和 95.23%，优于现有方法。此外，该方法在处理噪声数据和动态手势方面稳定性更好。

关键词 手势识别；特征提取；可变形卷积；骨架序列；全局信息

中图分类号 TP391.41 **文献标志码** A DOI 10.12178/1001-0548.2022401

Global Spatio-Temporal Deformable Network for Skeleton-Based Gesture Recognition

SHI Dongzi[1], LIN Honghui[1], LIU Yijiang[1], and ZHANG Xin[1, 2*]

(1. School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China;
2. Guangdong Artificial Intelligence and Digital Economy Laboratory, Guangzhou 510640, China)

Abstract The key of gesture recognition based on skeleton sequence is how to fuse spatio-temporal information and extract discriminate features. This paper proposes a key point focusing module. Through the global context modeling and the convolution method not limited to the fixed form, the network can span multiple frames and irrelevant key points, adaptively aggregate key point information closely related to gesture actions in the global scope, and extract the spatio-temporal characteristics of gesture. Experiments on Chalearn2013 and SHREC datasets show that the accuracy of our proposed method can reach 94.88% and 95.23%, and the method outperforms state-of-the-art methods. In addition, the method has better stability in dealing with noisy data and dynamic gestures.

Key words gesture recognition; features extraction; deformable convolution; skeleton sequence; global information

手势是一种符合人际交流习惯的表达方式，在智能家居、智能驾驶、体感游戏等领域得到了广泛的应用。一般而言，手势可以由多种模态表示，如彩色（RGB）视频流、深度视频流、光流、人体骨架序列等。其中，人体骨架序列描述的是预定义的人体关键点在手势运动过程的轨迹，为手势识别传递了重要的信息。随着光电技术和人体姿态估计算法的发展，骨架数据可以通过深度传感器直接获取，或者通过 RGB 图像进行关节估计[1-2]。

当用人体骨架关键节点的坐标变化进行手势描述时，手势的相对变化关系较为明显，有助于计算机更好地理解一些较为复杂的动作，从而提高手势识别的准确率。与基于 RGB 图像数据的方法相比，基于骨架数据的手势识别在面对遮挡、嘈杂背景、相机视角变化和照明变化等表现出优异的性能。然而，传统基于骨架数据的方法受限于手工设置的特征提取模式，网络表达受限。因此，本文提出全局时空可变形网络，其中关键点聚焦模块可以在全局范围内自适应地组合与手势动作密切相关的关键点，提取时空特征。

骨架序列可以视为一种关节坐标序列的时空矩阵。以拍手动作为例，与该手势相关的左手腕、左

收稿日期：2022−11−22；修回日期：2023−03−01
基金项目：中央高校基本科研业务费交叉学科研究项目（2022ZYGXZR104）；广东省数字孪生人重点实验室项目（2022B1212010004）
作者简介：石东子，主要从事手势识别方面的研究。
*通信作者 E-mail: eexinzhang@scut.edu.cn

手和右手节点应该关联在一起，并且得到每一帧手部的空间距离作为特征来描述这种手势。传统卷积神经网络（Convolution Neural Network, CNN）[3-4]可以自然地融合时空信息，但受限于固定的卷积形式。网络往往利用标准 3×3 卷积直接聚合相邻关键点信息，而与手势相关的关键点有可能在时空矩阵上不相邻，如左手和右手。网络只能通过堆叠局部卷积操作关联这些关键点。文献 [5] 提出 D-Pose 组合一维卷积和长短时记忆网络（Long Short Term Memory, LSTM）构造了 Conv-LSTM 结构来分别构建可学习的空间连接并提取动态信息。但上述过程只能单独在时间或空间上直接联系相关节点，无法交叉时间和空间信息进行交错卷积，网络的表达能力受限。

本文提出的关键点聚焦模块将卷积进行一定程度的空间偏移，聚合手势序列在时空维度上不相邻的关键信息。

由于骨架序列时空矩阵的时间维度往往比空间维度更长，同时，考虑到不同手势的复杂程度不同、手势执行速度不同、开始和结束的时间不一致等，时间维度上的远程建模更有挑战性。采用循环神经网络（Recurrent Neural Network, RNN）[6-7]可以捕获远程时间动态，但其计算复杂，难以满足实际需求，并且难以同时融合时空信息。因此，关键点聚焦模块会计算捕捉时域远程依赖的全局信息，使卷积在进行空间偏移时考虑到手势具有不同的时长，每个手势的开始、结束和执行时间并不一致。

综上，全局时空可变形网络能够关联关键点并学习手势的时间动态信息，这对于特征提取和手势识别至关重要。

1　相关工作

1.1　基于骨架的动作识别

按照网络结构划分，基于骨架的动作识别可以分为基于 CNN、基于 RNN 和基于 GCN 的方法[8-11]。CNN 聚合骨架时空矩阵信息，RNN 结构则适合对骨架时间序列数据进行建模，但由于其缺乏空间建模能力，文献 [5] 将 CNN 和 RNN 结合分别进行骨架的时空建模。另外，由于图结构能自然表示人体的结构和连接，基于 GCN 的方法也受到广泛关注，但图卷积只能从固定邻接矩阵提取特征，网络的表达能力同样受限。近年来，随着 Transformer 的兴起，也有文献引入 Attention 机制进行时空建模[11]。

对于骨架数据的时空特征提取，之前的工作是在时域和空间域上各自建模之后进行融合，文献 [9] 分别构建注意力机制空间图和时间维度的固定图，文献 [10] 则分别在时空维度上计算帧和关键点动作识别的重要性并进行时空交替训练，文献 [12] 则分别在时空域利用图卷积提取特征。利用 CNN 处理输入的骨架时空矩阵信息可以自然融合时空信息，本文提出了关键点聚焦模块提取时空特征，希望可以自适应组合不同帧上与手势相关的关键点信息。

1.2　上下文建模

卷积运算只能处理一个局部的区域，文献 [13] 提出空洞卷积，在不增加额外计算量的情况下扩大感受野，利用不同尺寸的卷积核提取多尺度特征，但需要重复应用卷积操作才能捕捉远程依赖。为了获得全局感受野，NL Net 结合了自注意力机制，通过计算任意两个位置之间的交互直接捕捉远程依赖[14-15]，但网络计算量较大。SENet 则对不同通道进行缩放得到全局特征并学习各个通道的权重系数从而建模通道与全局的依赖关系[16]，但对于全局上下文建模不够有效。Global Context Net 结合了 NL Net 和 SENet 的优点，使网络有效进行全局上下文建模的同时不引入过多计算参数[17]。

全局上下文建模已被证明对于图像识别、对象分割、动作定位[18] 等方面非常有用，本文主要探讨手势动作在全局范围内的特征提取以帮助网络理解不同时间依赖性的手势。

2　全局时空可变形网络

2.1　可变形卷积

一般的 2D 标准卷积可以分为两个步骤：1）用规则网格G在输入特征图x上采样；2）将经过$\boldsymbol{w}$加权的采样值相加。G定义了感受野的大小，当卷积核大小为$K\times K$，其中$K=2\times k+1$时，网格G可以表示为：$G=\{(-k,-k),(-k,-(k-1)),\cdots,(k,(k-1)),(k,k)\}$，故经过 2D 卷积输出特征 $\boldsymbol{y}$ 上p_0点可以表示为：

$$\boldsymbol{y}(p_0)=\sum_{p\in G}\boldsymbol{w}(p)\cdot\boldsymbol{x}(p_0+p) \tag{1}$$

式中，p为规则网格G上的点的枚举。

由于网格G中的规格网格使标准卷积难以适应复杂的几何形变，具体到骨架序列数据来说，当手势动作涉及多个帧时，时间距离较远的关键节点可

能无法较好地关联在一起，进而无法提取到较为准确的特征。文献 [19] 通过对每个采样点的位置添加一个 2D 的偏移变量，卷积核可以在原先位置的附近进行卷积，不再局限于之前的规则网格，时空融合特征有了更好的表达。2D 偏移变量通过一个标准卷积计算可得，则经过偏移，输出特征图 $\boldsymbol{y}$ 上 p_0 点可以表示为：

$$\boldsymbol{y}(p_0)=\sum_{p\in G}\boldsymbol{w}(p)\cdot\boldsymbol{x}(p_0+p+\Delta p_n) \tag{2}$$

加入卷积核的空间偏移 Δp_n 之后，网络可以更直观地识别手势关键节点及其随时间变化的情况，如拍手动作，卷积可以聚合左右手和手腕节点的信息，因此能更准确地表征某一特定手势。

可变形卷积具体通过一个 3×3 卷积来学习相应卷积核的空间偏移 Δp_n。由于手势动作的持续时间存在差异，通过标准卷积学习到的空间偏移存在一定的局限性。为了捕捉远程依赖，学习多尺度时间信息，本文在可变形卷积基础上加入全局信息，提出了关键点聚焦模块。

2.2 关键点聚焦模块

在进行可变形卷积之前，本文先获取全局信息特征，如图 1 所示。将输入骨架特征用一维序列表示，定义为 $\boldsymbol{x}=\{x_i\}_{i=1}^{N_p}$，其中 N_p 为特征的位置数目，即骨架序列帧数和关节点数目之积。全局信息特征 $\boldsymbol{z}=\{z_i\}_{i=1}^{N_p}$ 可定义为[15]：

$$z_i=x_i+\frac{\boldsymbol{W}_z}{C(x)}\sum_{j=1}^{N_p}f(x_i,x_j)g(x_j) \tag{3}$$

式中，$g(x_j)$ 计算了不同位置 j 输入的特征映射，为了简单，本文使用了线性嵌入，令 $g(x_j)=\boldsymbol{W}_v x_j$，$\boldsymbol{W}_v$ 为需要学习的权重矩阵；$\boldsymbol{W}_z$ 表示线性变换矩阵，当 $\boldsymbol{W}_z$ 为 0 时，输入即为 x_i。

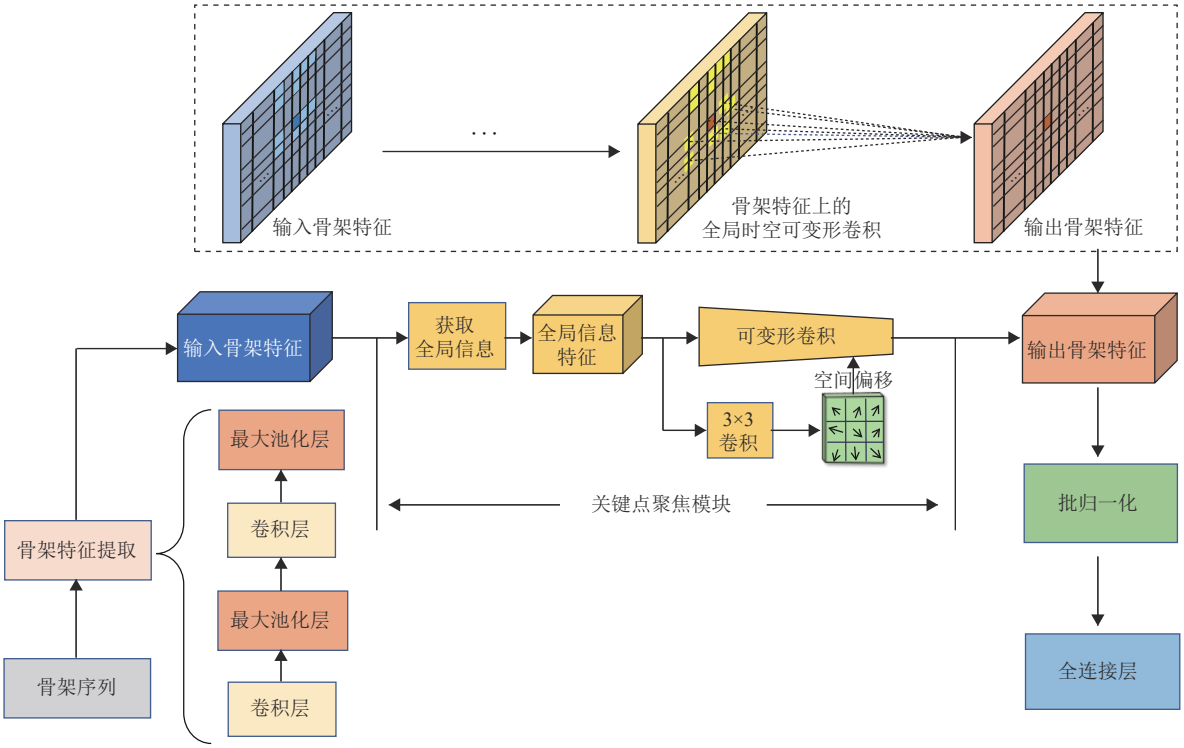

图 1 全局时空可变形网络结构

$f(x_i,x_j)=e^{q(x_i)^{\mathrm{T}}k(x_j)}$ 表示 i 和 j 之间的相似关系，这里 $q(x_i)=\boldsymbol{W}_q x_i$，$k(x_j)=\boldsymbol{W}_k x_j$。本文设 $C(x)$ 为归一化响应，则归一化的相似关系表示为：

$$\frac{f(x_i,x_j)}{C(x)}=\frac{\exp(<\boldsymbol{W}_q x_i,\boldsymbol{W}_k x_j>)}{\sum_{m=1}^{N_p}\exp(<\boldsymbol{W}_q x_i,\boldsymbol{W}_k x_m>)} \tag{4}$$

由于 z 不受位置依赖，且 $\boldsymbol{W}_z$ 对结果影响不大[17]，本文可以通过计算全局注意力图对式 (3) 进行简化，最终输出的全局信息特征可以表示为：

$$z_i=x_i+\boldsymbol{W}_v\sum_{j=1}^{N_p}\frac{\exp(\boldsymbol{W}_k x_j)}{\sum_{m=1}^{N_p}\exp(\boldsymbol{W}_k x_m)}x_j \tag{5}$$

获取全局信息特征如图 2 表示，骨架特征经过特征映射后通过 softmax 计算出注意力分数，再和原始骨架特征相乘累加，之后再经过一次映射，并加上原始骨架特征得到全局信息特征。全局时空可变形网络加上原始骨架特征有利于保障模型的性能，并在式 (5) 基础上引入瓶颈设计，在 1×1 卷积时减少通道以降低参数量。全局时空可变形网络还加入 LayerNorm 和 ReLu 帮助模型收敛。

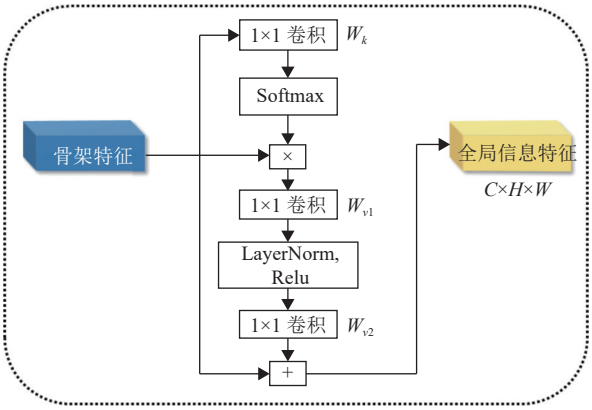

图 2　全局信息特征的获取

在获取全局信息特征后，本文再利用 3×3 卷积学习全局空间偏移Δq_n，如图 1 所示。则输出特征$\boldsymbol{y}$上p_0点可以表示为：

$$\boldsymbol{y}(p_0)=\sum_{p\in G}\boldsymbol{w}(p)\boldsymbol{z}(p_0+p+\Delta q_n) \tag{6}$$

在提取骨架序列的全局上下文信息基础上，全局可变形网络学习整个时间序列上手势关键节点的变换，将手势持续时间的差异融入特征描述中。

2.3　全局时空可变形网络结构

在全局时空可变形网络中，骨架序列首先经过卷积层和最大池化层进行骨架特征的提取，如图 2 所示，最大池化层用于降低特征维度，卷积层由 3 个 3×3 标准卷积和可变形卷积组成。

文献 [19] 指出，多次可变形偏移可以产生一个叠加的效果。本文在网络深处引入关键点聚焦模块，网络此时已经过多层卷积聚合信息，有较大的感受野，有助于更好地关联时间尺度长且幅度较大的手势的相关节点。关键点聚焦模块和普通的卷积模块有着相同的输入和输出。因此，本文直接将它取代普通的卷积。

在经过关键点聚焦模块之后，网络学习到了对识别手势更有针对性的特征，通过批归一化层，帮助调整特征的分布，加快模型的收敛速度，最后经过全连接层，预测每一个类别的概率，得到分类结果。

3　实验与结果分析

3.1　数据集

本文在 ChaLearn2013[20] 和 SHREC 数据集[21] 上评估了本文提出的方法。

ChaLearn2013 多模态手势数据集，包含了 27 个不同的人执行的 20 个意大利手势。该数据集提供 RGB、深度、前景分割和 Kinect 骨架，分为训练集、验证集和测试集，分别包含 6 850、3 454 和 3 579 个样本。本文只使用骨架数据进行手势识别。

SHREC 数据集包含由 28 名参与者（所有参与者均为右手）执行的 14 个动态手势，并由 Intel RealSense 近程深度相机捕捉。每一个手势由每个参与者以两种方式执行 1～10 次：使用一个手指和整只手。数据集由捕获的 2 800 个序列组成。对数据集中每个序列的每个帧，将保存分辨率为 640×480 的深度图像和 22 个关节的坐标（在二维深度图像空间和三维世界空间中）。本文中仅使用骨架数据。

3.2　实验设置

本文通过随机梯度下降法对网络进行参数更新，并将批次大小设置为 32，学习率在 3×10^{-6}～9×10^{-3} 之间更新，步长为 1 060。该网络在 Pytorch 框架中实现，使用 GeForce GTX 1080 GPU 进行训练。

3.3　消融实验

3.3.1　噪声

由于手部区域较小，背景干扰严重，指尖运动自由度高，交互过程存在自遮挡等因素，手部姿态估计是一个较难任务。而基于手部姿态估计算法获得的骨架序列不可避免存在一定噪声干扰。为验证全局时空可变形网络对噪声的鲁棒性，本文在 ChaLearn2013 上进行了关于噪声的消融实验，以证明本文提出的方法对噪声的鲁棒性。本文向关键点随机添加最大振幅为 0.1 的随机噪声，并以标准卷积为参照进行对比，结果如图 3 所示。从图中可以看出噪声的百分比增强的情况下，网络的准确率保持在较高的平稳水平，噪声的鲁棒性较好。

3.3.2　多尺度

正如上述所说，对于不同时间尺度的手势动作，本文需要捕获不同时间依赖性的信息。但是因为手势多样性和个体之间的差异，网络很难捕捉到具有代表性的特征。对此本文进行了实验，首先以

固定的采样率（如 39）对网络进行训练，之后改变骨架数据的采样率，如 30、35、45 和 50，来模拟不同时间尺度的手势，结果如图 4 所示，本文提出的关键点聚焦模块可以充分捕获多尺度时间信息，大幅优于基线。

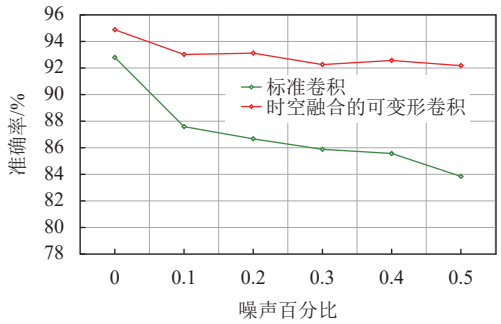

图 3　噪声对准确率的影响

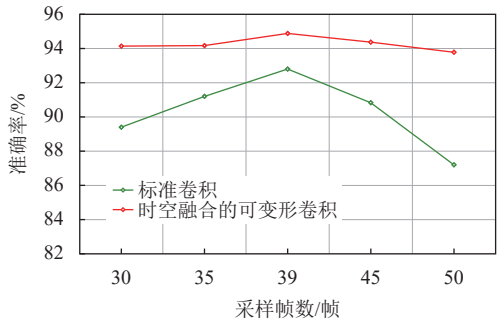

图 4　时间尺度对准确率的影响

3.4　关键点聚焦模块可视化

图 5 为关键点聚焦模块的可视化，上面两行表示同一动作“What I would do to you”的两个不同手势序列，第三行是动作“Enough”的手势序列，颜色代表关键节点的注意力分数，由蓝到红注意力分数变高。其中，骨架序列分别是第 0、4、8、16、20、24、29 和 34 时的骨架姿态。从图 5 前两行的骨架序列图可以看出，虽然是相同的动作，但是第一行的骨架序列显示 T=0 时就开始执行动作，到 T=29 动作基本结束，而第二行的骨架序列显示 T=8 之后才开始执行动作，到 T=34 基本结束。全局时空可变形网络关注的是执行动作期间关节点变化情况，因此在对应的执行时间内，关节点的颜色表示网络的不同关注程度。

而对比前两行的动作和第三行的动作可以看出，在空间维度上，网络对不同动作的关节点关注情况也不同。前两行用到了双手表示动作，网络关注到了两边手、手腕和肩膀的关节点；而第三行主要是用到了单手表示，因此注意力基本只集中在了执行动作的手腕和手肘上。

3.5　与先进方法的比较

本文在 ChaLearn2013 数据集上进行训练。本文将全局时空可变形网络与其他先进算法进行了比较，结果如表 1 所示。

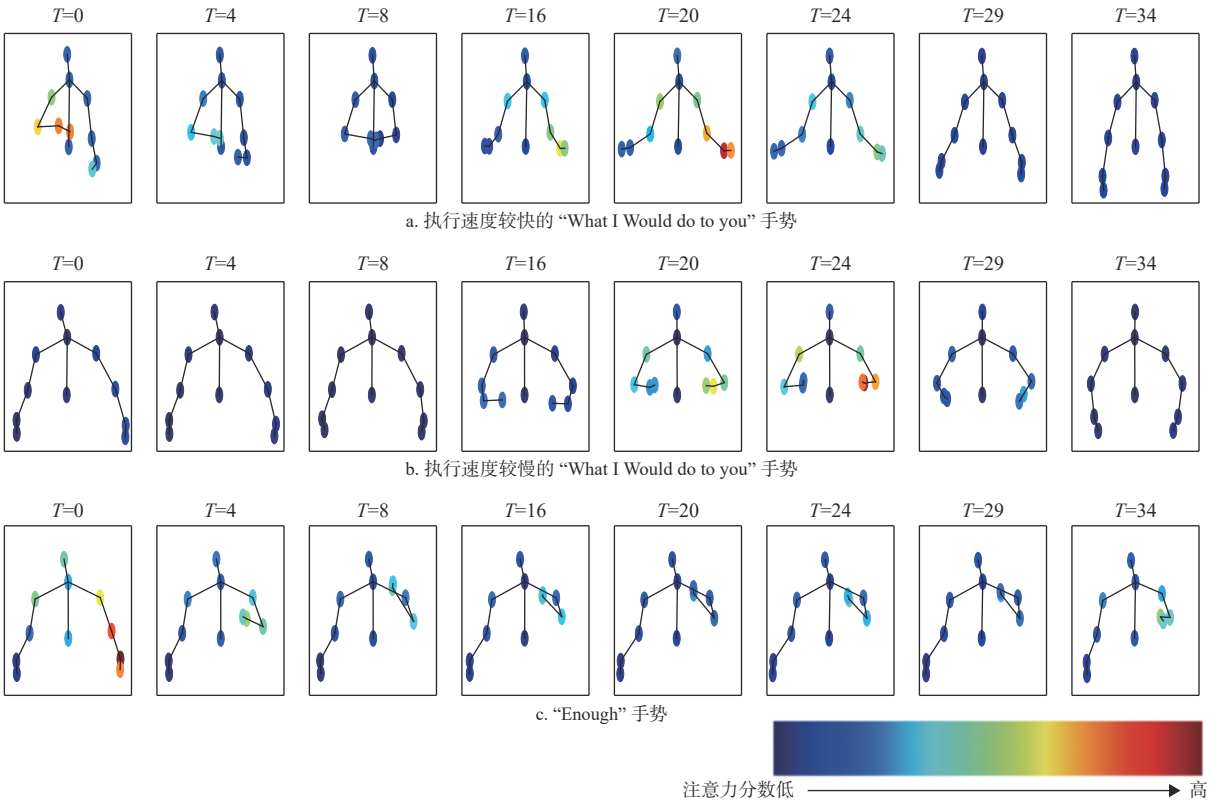

图 5　骨架序列及全局注意力可视化

表 1　ChaLearn2013 数据集上不同方法对比

网络架构	方法	准确率/%
基于RNN	PT-Logsig-RNN[8]	93.27
	Two-stream RNN[9]	91.70
CNN和LSTM	D-Pose [5]	92.54
基于GCN	CTR-GCN[22]	92.82
	GCN-Logsig-RNN[23]	92.86
基于Transformer	STFFormer[24]	92.08
	ST-TR[25]	93.50
基于CNN	Multi-path CNN [3]	93.13
	CNN for Skeleton [4]	91.16
	全局时空可变形网络（本文）	**94.88**

文献 [8-9] 采用 RNN 架构进行远程时间建模，空间上的建模能力不足。文献 [5] 同样使用可变形卷积关联空间上相距较远的点，但是无法交错时空上相关的点。文献 [22-23] 虽然有骨架的连接图作为预定义的知识，但是动作可能并不关联到相连关节，如拍手动作左右手节点应该较为相关，图卷积不一定能够关注到。文献 [24-25]采用全局感受野设计，但只引入全局信息可能无法捕捉到关键节点信息。文献 [3-4] 虽然通过设计长方形卷积核和不同骨架模态来丰富网络的特征表达，但还是受 CNN 的固定卷积核的影响，不能跨越关联关键节点。

这些方法主要由 RNN、GCN、Transformer 和 CNN 等模型改进发展而来。由于关键点聚焦模块对于时空特征灵活融合和全局信息的考虑，本文的方法取得了最佳的结果。

在 SHREC 数据集上进行对比的结果如表 2 所示。文献 [26] 采用多尺度掩码注意力机制进行全局的学习，但掩码可能没有注意到手势动态变化关键的节点；文献 [27] 则着重捕捉手势的动态变化；文献 [28] 关注到了手势姿态的变化和移动以此进行时域和空域的建模。在该数据集中，各个关节点的位置距离相近，在做手势动作的时候，关节点之间的变化幅度相对较小，这些方法结合时空建模的能力可能还有所不足。本文认为关节点之间的本身相关程度比较高，可以较容易提取到相对有用的全局特征，因此本文将网络中前两个可变形卷积模块全都替换为关键点聚焦模块，最终可以达到 95.23% 的准确率，相比起其他方法有所提升。

表 2　SHREC 数据集上不同方法对比

方法	准确率/%
STA-Res-TCN[26]	93.6
TCN-Summ[27]	93.57
HPEV[28]	94.88
本文	**95.23**

4　结束语

本文主要基于骨架数据提出了关键点聚焦模块。通过可变形卷积将时空信息交错以使关键帧和关键节点可以得到有效关联，以此提取手势特征。加入全局信息模块，让网络可以根据手势复杂度进行动态调整，学习更有意义的可变形的空间偏移。在 ChaLearn2013 和 SHREC 数据集两个数据集上的实验结果表明，本文方法在精确度上优于已有的方法。此外，额外的实验表明，本文提出的模块在处理不同时间尺度的噪声数据和动态手势时更具鲁棒性。

本文在进行时空建模时，主要还是利用 CNN 逐渐增大感受野，当关键点特别多时，可能没办法很好地关联在一起，但 Transformer 直接提取丰富的全局信息，有更强的时空建模能力。未来可以探索如何在这些全局信息中进行筛选，从而锁定和动作相关的关键点信息。同时，本文只利用骨架数据的三维空间坐标信息，未来可以探索多种模态融合，通过交换不同模态的知识提高识别的准确率。

参 考 文 献

[1] GE L H, CAI Y J, WENG J W, et al. Hand pointNet: 3D hand pose estimation using point sets[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8417-8426.

[2] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7291-7299.

[3] LIAO L F, ZHANG X, LI C Y. Multi-path convolutional neural network based on rectangular kernel with path signature features for gesture recognition[C]//2019 IEEE Visual Communications and Image Processing (VCIP). Sydney: IEEE, 2019: 1-4.

[4] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network[C]//2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur: IEEE, 2015: 579-583.

[5] WENG J W, LIU M Y, JIANG X D, et al. Deformable pose traversal convolution for 3D action and gesture

recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 136-152.

[6] WANG H S, WANG L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 499-508.

[7] LIAO S J, LYONS T, YANG W X, et al. Learning stochastic differential equations using RNN with log signature features[EB/OL]. (2019-08-22)[2022-11-20]. https://arxiv.org/pdf/1908.08286.pdf.

[8] YAN S, XIONG Y J, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//The 32nd AAAI Conference on Artificial Intelligence. New Orleans Louisiana: AAAI Press, 2018.

[9] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12026-12035.

[10] LI M S, CHEN S H, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3595-3603.

[11] 石跃祥, 朱茂清. 基于骨架动作识别的协作卷积 Transformer 网络[J]. 电子与信息学报, 2022, 44: 1-9.
SHI Y X, ZHU M Q. Collaborative convolutional transformer network based on skeleton action recognition[J]. Journal of Electronic and Information Technology, 2022, 44: 1-9.

[12] 李扬志, 袁家政, 刘宏哲. 基于时空注意力图卷积网络模型的人体骨架动作识别算法[J]. 计算机应用, 2021, 41(7): 1915-1921.
LI Y Z, YUAN J Z, LIU H Z. Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model[J]. Journal of Computer Applications, 2021, 41(7): 1915-1921.

[13] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-06-17)[2022-11-20]. https://arxiv.org/pdf/1706.05587.pdf.

[14] WANG F, WANG G R, HUANG Y W, et al. Sast: Learning semantic action-aware spatial-temporal features for efficient action recognition[J]. IEEE Access, 2019, 7: 164876-164886.

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

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[17] CAO Y, XU J, LIN S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul: IEEE, 2019.

[18] CHEN P, GAN C, SHEN G, et al. Relation attention for temporal action localization[J]. IEEE Transactions on Multimedia, 2019, 22(10): 2723-2733.

[19] DAI J, QI H, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Honolulu: IEEE, 2017: 764-773.

[20] ESCALERA S, GONZÀLEZ J, BARÓ X, et al. Multi-modal gesture recognition challenge 2013: Dataset and results[C]//Proceedings of the 15th ACM on International Conference on Multimodal Interaction. New York: Association for Computing Machinery, 2013: 445-452.

[21] SMEDT Q, WANNOUS H, VANDEBORRE J P, et al. 3D hand gesture recognition using a depth and skeletal dataset[EB/OL]. [2022-10-24]. https://doi.org/10.2312/3dor.20171049.

[22] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 13359-13368.

[23] LIAO S J, LYONS T, YANG W X, et al, Logsig-RNN: A novel network for robust and efficient skeleton-based action recognition[EB/OL]. [2022-10-25]. https://arxiv.org/pdf/2110.13008.pdf.

[24] QIU H L, HOU B, REN B, et al. Spatio-temporal tuples transformer for skeleton-based action recognition[EB/OL]. [2022-11-08]. https://arxiv.org/pdf/2201.02849.pdf.

[25] PLIZZARI C, CANNICI M, MATTEUCCI M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding, 2021, 208: 103219.

[26] HOU J X, WANG G J, CHEN X H, et al. Spatial-temporal attention Res-TCN for skeleton-based dynamic hand gesture recognition[C]//European Conference on Computer Vision. Munich: Springer, 2018.

[27] SABATER A, ALONSO I, MONTESANO L, et al. Domain and view-point agnostic hand action recognition[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 7823-7830.

[28] LIU J B, LIU Y C, WANG Y, et al. Decoupled representation learning for skeleton-based gesture recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 5751-5760.

编 辑 张 莉