控制与决策
Control and Decision

基于注意力引导空域图卷积SRU的动态手势识别

陈炫琦, 佘青山, 张波涛, 马玉良, 张建海

引用本文:
陈炫琦,佘青山,张波涛,马玉良,张建海. 基于注意力引导空域图卷积SRU的动态手势识别[J]. 控制与决策, 2023, 38(11): 3083−3092.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2073

基于注意力引导空域图卷积SRU的动态手势识别

陈炫琦[1], 佘青山[1†], 张波涛[1], 马玉良[1], 张建海[2]

(1. 杭州电子科技大学 自动化学院，杭州 310018；2. 浙江省脑机协同智能重点实验室，杭州 310018)

摘 要: 基于手部骨骼的动态手势识别是计算机视觉和人机交互领域的一个研究热点. 手势涉及的关节在空间上分布更紧密,相关性更强. 针对目前基于骨骼的动态手势识别存在空间特征复杂、识别计算速率缓慢等问题,提出一种注意力引导空域图卷积简单循环单元(ASGC-SRU)网络. 首先,将空域图卷积嵌入至SRU的门结构中,使得具有高速并行计算能力的SRU能够对复杂手势的时域和空域信息进行建模;然后,引入一种指关节注意力引导模块,使得更重要的指关节具有更高的关注度;最后,引入一种注意力增强空域图丢弃(ASD)的正则化方法,缓解网络过拟合的弊端. 为验证所提出方法的有效性,在公认的动态手势数据集SHREC'17和DHG 14/28上进行大量实验,实验结果表明,所提出方法取得了较高的识别准确率,同时保持优良的计算效率.

关键词: 动态手势识别；图卷积；简单循环单元；注意力引导；过拟合

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2073

引用格式: 陈炫琦, 佘青山, 张波涛, 等. 基于注意力引导空域图卷积SRU的动态手势识别[J]. 控制与决策, 2023, 38(11): 3083-3092.

Dynamic gesture recognition based on attention-guided spatial graph convolutional SRU

CHEN Xuan-qi[1], *SHE Qing-shan*[1†], *ZHANG Bo-tao*[1], *MA Yu-liang*[1], *ZHANG Jian-hai*[2]

(1. College of Automation，Hangzhou Dianzi University，Hangzhou 310018，China；2. Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province，Hangzhou 310018，China)

Abstract: Dynamic gesture recognition based on hand skeleton is a research hotspot in the field of computer vision and human-computer interaction. The joints involved in gestures are more closely distributed in space and have stronger correlation. Aiming at the problems of complex spatial features and slow recognition calculation speed in skeleton-based dynamic gesture recognition, an attention-guided spatial graph convolution simple recurrent units (ASGC-SRU) network is proposed. Firstly, the spatial graph convolution is embedded into the gate structure of the SRU, so that the SRU with high-speed parallel computing ability can model the temporal and spatial information of complex gestures. Then, an attention-guided module of finger joints is introduced to make more important finger joints have higher attention. Finally, an attention-enhanced spatial dropgraph (ASD) regularization is introduced to alleviate the effect of the over-fitting problem. To verify the effectiveness of the proposed method, a series of experiments are conducted on the recognized dynamic gesture datasets SHREC'17 and DHG 14/28. The results show that the proposed method has achieved high recognition accuracy and simultaneously maintains excellent calculation efficiency.

Keywords: dynamic gesture recognition；graph convolution；simple recurrent unit；attention-guided；over-fitting

0 引 言

目前,手势识别作为一种直观且有效的交互工具,在许多交互体验中发挥着重要作用,应用于手语翻译[1]、虚拟现实[2]、辅助生活[3]等场景. 动态手势识别从序列数据中识别手势,具有更高的实用性. 现有的动态手势识别方法根据输入类型主要分为两类:基于图像的方法和基于手部骨架[4]的方法. 其中:基于图像的方法以RGB或RGB-D图像序列作为输入,但是这些序列易受到光照、背景以及遮挡等因素的影响. 基于手部骨架的方法减轻了背景混乱带来的

收稿日期: 2021-11-27；**录用日期:** 2022-06-24.

基金项目: 国家自然科学基金项目(61871427)；浙江省重点研发计划项目(2019C04018)；浙江省自然科学基金重点项目(LZ22F010003)；国家重点研发计划项目(2017YFE0116800).

责任编委: 谢晖.

†通讯作者. E-mail: qsshe@hdu.edu.cn.

困难,且具有较低的计算成本,因此能够更好地在移动设备上实现实时手势交互. 如今,随着如微软Kinect和英特尔RealSense这样具有高性价比的深度传感器的快速发展以及高精度姿态估计算法的出现,可有效地获得高精度的骨骼数据,因此基于骨架的动态手势识别成为一个活跃的研究领域.

近年来,传统的方法,如文献[5]从骨骼序列中人工提取特征,并将其提供给分类器进行手势识别,但是这些人工提取的特征并不能有效地模拟空间和时间信息. 文献[6]和文献[7]将手势骨架数据建模为一系列向量,再由卷积神经网络(convolutional neural network, CNN)或循环神经网络(recurrent neural network, RNN)进行处理,直接学习手部时空特征,但是忽略了手部关节间的运动学依赖关系. 时空图卷积神经网络[8]已成功应用于基于骨骼的人体动作识别. 文献[9]将图卷积嵌入长短期记忆网络(long short-term memory, LSTM)进行时域和空域的建模,用于人体动作识别,但是具有较高的计算复杂度,不适合具有高实时性的动态手势识别. 文献[10]和文献[11]将具有高速并行计算能力的简单循环单元(simple recurrent unit, SRU)应用于人体动作识别,具有更快的计算效率,但是缺少对骨架间结构相关性的建模.

本文对手势图结构序列进行时空建模,更好地利用指关节的结构性特征和运动特征,并保持优秀的计算识别效率,本文内容如下.

1) 提出一种基于注意力引导空域图卷积简单循环单元(attention-guided spatial graph convolutional-simple recurrent unit, ASGC-SRU)的动态手势识别模型. 将具有高速并行运算能力的SRU与空域图卷积(SGC)相结合,在动态手势识别中对时空特征信息进行建模.

2) 引入一种注意力引导机制,将指关节的重要性程度进行划分,使得网络对贡献更大、时空特征信息更有效的指关节给予更高的关注度.

3) 引入一种注意力增强的空域图丢弃(attention-enhanced spatial dropgraph, ASD)的正则化方法,以缓解对复杂动态手势训练过程中图卷积的过拟合问题.

1 相关工作

随着硬件设备的迅速发展,深度神经网络学习受到各研究领域的关注. 同样,在手势识别领域中以端到端的深度神经网络取代传统的人工特征提取算法,已成为目前主流的发展趋势.

Núñez等[12]提出了一种使用CNN提取每帧特征,并使用LSTM将CNN的输出聚合起来,但是难以解析特征的结构信息,且网络推理速度缓慢. Hou等[13]设计了一种端到端时空注意残差时间卷积网络(spatial-temporal attention residual temporal convolutional network, STA-Res-TCN),该网络对时间卷积网络[14]进行了修改,用于识别基于骨架的动态手势,然而该方法将X、Y、Z轴的三维坐标堆积在同一通道内,限制了其性能. Li等[15]提出了一种手势图卷积网络,在图中添加3种类型的边以精细地描述关节的连接作用,并扩展关节坐标的维数,但是缺乏不同关节间的重要性体现. 缪永伟等[16]将动态手势的运动分为手在空间的全局运动和手指在手内的局部运动两部分,并利用关键帧提取以解决时域信息处理问题,虽然具有较高的精度,但是难以避免人工特征提取和关键帧提取的繁琐成本花费. Liu等[17]将手势解耦为手部姿态变化和手部运动,姿态流通过设计的3D CNN融合了手部姿态的空间布局与时间信息,运动流通过2D CNN提取手部运动特征,从而对精细手势具有较强识别能力,但是需要对双流网络进行联合评估,大大增加了计算时间.

随着图卷积网络在人体动作识别中的广泛应用,可知图卷积对人体骨架这样的图结构数据具有强大的拟合能力. 本文在保留图卷积强大空间特征提取能力的同时,将其与在时间维度上对手部运动推理更具快速性的简单循环单元SRU相结合. 同时,在几乎不提升计算成本的基础上,引入轻量级的注意力引导模块、注意力增强的空域图丢弃机制,进一步优化网络对手势局部特征的推理能力.

2 注意力引导空域图卷积SRU

基于骨骼的动态手势识别对手部骨骼序列进行时空域建模. 本文设计了基于注意力引导空域图卷积SRU的端到端网络结构(ASGC-SRU),并引入注意力增强的空域图丢弃机制(attention-enhanced spatial dropgraph, ASD),兼顾良好的识别精度和速度.

本文所建模型的总体框架如图1所示,以多帧动态手势骨骼序列数据作为输入. 首先,对数据进行特征融合增强,将各骨骼节点的三维坐标通过全连接层映射至高维特征空间,再与骨骼运动流数据[18](节点帧间差异)进行特征拼接,丰富数据的特征信息;然后,通过4层ASGC-SRU层进行时空特征的建模,每层前采用平均池化,在时间维度上减少时间感受野,加快拟合速度. 在ASGC-SRU层中嵌入空域图卷积算子,更好地捕捉手骨骼数据的单帧空域静态特征和多帧时域动态特征. 在其中引入注意力引导模块,

引导注意力关注于更重要的节点. 在图卷积算子中使用ASD策略,减小Dropout在手骨骼图中随机节点丢弃的弊端,使得更冗余的节点得以丢弃;最后,将网络单元末端隐藏状态输出与注意力权重分数加权后的隐藏状态分别通过全连接层后的特征进行聚合,以Softmax函数实现最终的手势预测.

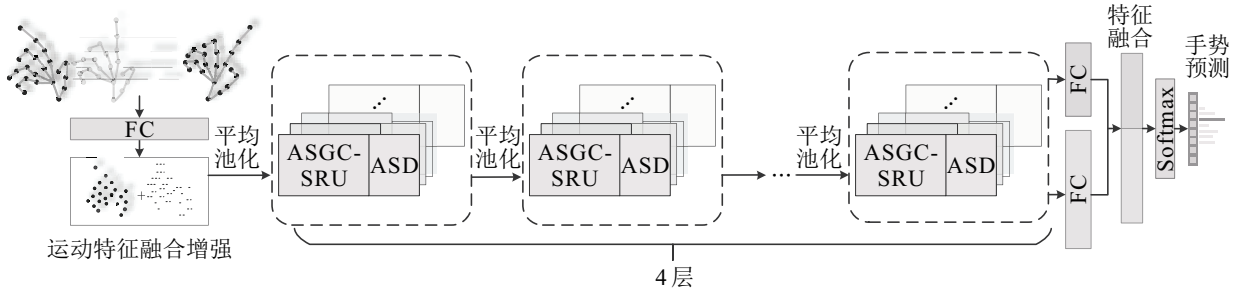


图 1　模型总体框架

2.1　空域图卷积SRU

2.1.1　图结构初始化

动态手势骨骼序列通常由每帧中手关节的二维或三维坐标表示. 在空域上,对具有 N 个关节和 t 帧的手骨骼序列,构造一个无向空间图 $G_t = (V_t, E_t)$,表示第 t 帧时手部的骨架图. 其中:节点集 $V_t = \{v_{t_1}, v_{t_2}, \ldots, v_{t_N}\}$ 包括了整个手部骨骼中的所有关节,节点集 $E_t = \{(v_{t_i}, v_{t_j}) | v_{t_i}, v_{t_j} \in V_t\}$ 为手部骨骼关节点间的无向连接,即骨架边的集合. 同时,本文将节点 v_{t_i} 的邻接节点集定义为 $\alpha(v_{t_i}) = \{v_{t_j} | d(v_{t_i}, v_{t_j}) \leqslant 1\}$, $d(v_{t_i}, v_{t_j})$ 为从节点 v_{t_i} 到 v_{t_j} 的最小路径,即采样距离,本文默认其均小于等于1.

2.1.2　手部空域图卷积

受ST-GCN[8]对人体骨架图划分的思想启发,本文对手部骨骼空域图的邻接节点集进行区域划分. 根据手部物理结构,如抓握、张开等,远离手掌心的关节总是由靠近手掌心一侧的相邻关节所控制,因此本文将掌心节点定义为图的根节点. 如图2所示,根据人体骨架图的经验,将邻接节点集划分为3个子集,分别为:1) 顶点本身(黑色);2) 向心集,更靠近掌心的相邻指关节(灰色);3) 离心集,远离掌心的相邻指关节(白色).

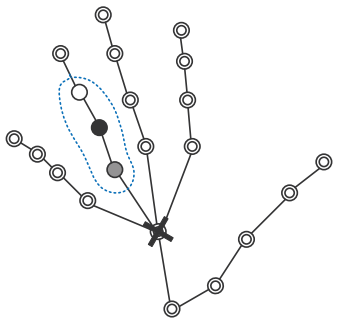


图 2　图卷积节点子集划分示意图

图2中:采用图卷积运算提取空域中手势图的高级特征,对于 t 时刻顶点 v_{t_i} 上的图卷积运算公式为

$$X_{\text{out}}(v_{t_i}) = \sum_{v_{t_j} \in \alpha(v_{t_i})} \frac{1}{Q(v_{t_j})} X_{\text{in}}(v_{t_j}) w(\varphi_{t_i}(v_{t_j})). \tag{1}$$

其中: v_{t_j} 为输入手部图的顶点; $X_{\text{in}}(v_{t_j})$、$X_{\text{out}}(v_{t_i})$ 分别为相邻顶点 v_{t_j}、v_{t_i} 的输入和输出特征; $w(\cdot)$ 类似于原始卷积运算中的权重函数,通过给定的输入提供一个权重向量; $\varphi_{t_i}(\cdot)$ 为一种图标记函数,将节点 v_{t_j} 的邻接节点集划分为数量为 $Q(v_{t_j})$ 的子集,使得每个子集均有与其对应的权重向量,子集数 $Q(v_{t_j})$ 等于3. 最终将特征参数除以子集数量,从而平衡各子集的影响程度.

在式(1)中,通过引入邻接矩阵并变换,可得到如下图卷积运算公式:

$$X_{\text{out}} = \sum_{q=1}^{Q} D_q^{-\frac{1}{2}} A_q D_q^{-\frac{1}{2}} X_{\text{in}} w_q. \tag{2}$$

其中: N 遵循上述的分区策略,设置为3,即3种不同的子集;通过式(1)的图标记函数 φ_{t_i} 得到子集的标签索引 $q \in \{1, 2, 3\}$; A_q 为不同标签索引的子集所对应的邻接矩阵,当对应的第 n 个子集中节点 v_{t_i} 与 v_{t_j} 存在无向连接时, $A_{i,j} = 1$,否则 $A_{i,j} = 0$,且存在节点自连接; D_q 为不同标签索引子集的归一化度矩阵,即 $D_q^{ii} = \sum_j (\overline{A}_q^{ij})$.

2.1.3　空域图卷积SRU

为了实时且快速地人机交互,并保证识别的精度,本文使用简单循环单元SRU[19]与空域图卷积相结合进行手势识别. 与RNN、LSTM等传统的循环递归网络单元相比,SRU在保证良好识别精度的前提下,大大提升了网络拟合速度. 因为SRU大大简化了细胞状态和隐藏状态的计算,使得隐藏状态不再依赖于上一时刻的隐藏状态.

虽然SRU具有更高的效率,但是动态手势中包含很多较为精细的手势动作,不仅与整只手的空间位移有关,且依赖于手指间的空间相关性. 于是,本文提出注意力引导的空域图卷积SRU(ASGC-SRU)模型,将空域图卷积运算取代SRU中的全连接运算,同时引入手骨骼关节注意力引导机制(见2.2节)以关注图中更重要的节点. 图3为ASGC-SRU结构.

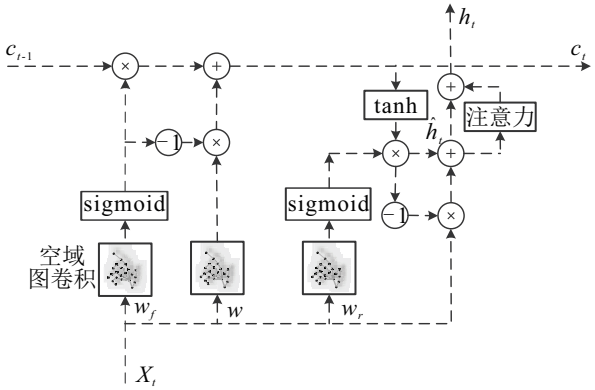


图3 ASGC-SRU结构

图卷积与SRU结合后具有时空特性,对动态手势进行时空域联合建模. 以手部骨骼图的形式输入网络,通过遗忘门和重置门,不断更新隐藏状态和细胞状态. ASGC-SRU模型的表达式为

$$\hat{X}_t = w(\gamma^*)X_t, \tag{3}$$

$$f_t = \text{sigmoid}(w_f(\gamma^*)X_t + b_f), \tag{4}$$

$$r_t = \text{sigmoid}(w_r(\gamma^*)X_t + b_r), \tag{5}$$

$$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot \hat{X}_t, \tag{6}$$

$$\hat{h}_t = r_t \odot \tanh(c_t) + (1 - r_t) \odot X_t, \tag{7}$$

$$h_t = f_a\hat{h}_t + \hat{h}_t. \tag{8}$$

其中:(γ^*)为空域图卷积算子;$\odot$为哈达玛乘积;X_t为t时刻的输入特征;w、w_f、w_r、b_f、b_r为在训练过程中可学习的权重参数和偏执量;f_t、r_t分别为遗忘门和重置门;c_t、$\hat{h}_t$分别为细胞状态和原始隐藏状态;h_t为通过注意力引导模块(见2.2节)的隐藏状态输出;f_a为一种注意力引导网络,能够自适应地将注意力集中于关键的手部关节,对于更为精细的手势分类具有更强的鲁棒性.

2.2 注意力引导模块

基于手骨骼的动态手势识别过程依赖于手关节的三维坐标,对于一些精细手势的识别,部分关节对整个手势动作起主导作用,而另一部分关节对手势的影响程度较低. 为了增强部分重要关节的关注程度,本文进一步设计了指关节注意力引导模块(attention-guided module),将其与空域图卷积SRU相结合形成ASGC-SRU层,能够自适应地将注意力向重要的指关节引导.

指关节注意力引导模块如图4所示,右侧为一种轻量的注意力引导网络. ASGC-SRU的隐藏状态中具有丰富的空域结构信息和时域动态信息,将单元末端的原始隐藏状态$\hat{h}_t$输入注意力引导网络. 其中:每个关节的特征信息通过全连接层与tanh激活函数后形成指关节聚合特征,再通过一个全连接层后,利用非线性激活函数sigmoid计算0～1间的注意力权重分数,计算如下式所示:

$$a_t = \text{sigmoid}\Big(w_2\tanh\Big(\sum_{i=1}^{N} w_1\hat{h}_{t_i} + b_1\Big) + b_2\Big). \tag{9}$$

其中:a_t为各节点的注意力权重分数,w_1、w_2为全连接层中可训练的权重参数,b_1、b_2为偏执量.

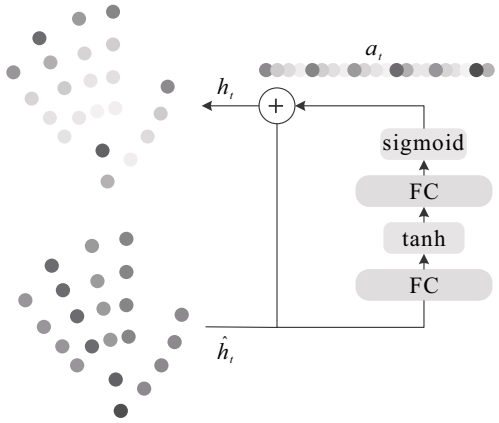


图4 指关节注意力引导模块

在最后一层ASGC-SRU层的末端,将关节隐藏状态输出h_t与注意力权重分数加权后的原始隐藏状态$a_t\hat{h}_t$分别通过全连接层后,对特征先进行拼接聚合,再将t时刻的聚合特征F_t转换为J个手势类别的分数$s_t \in \{s_{t_1}, s_{t_2}, \ldots, s_{t_J}\}$,最终采用softmax函数计算第$i$类的预测概率,有

$$\hat{Y}_{t_i} = \frac{\mathrm{e}^{s_{t_i}}}{\sum\limits_{j=1}^{J}\mathrm{e}^{s_{t_j}}},\ i = 1, 2, \ldots, J. \tag{10}$$

于是,引入指关节注意力引导模块后,本文采用的损失函数如下式所示:

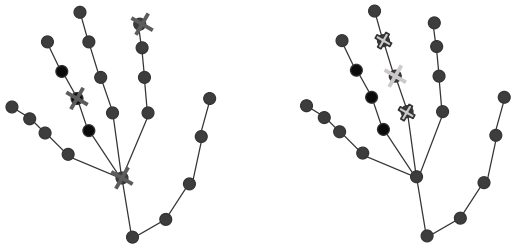
$$L = -\sum_{t=1}^{T_3}\sum_{i=1}^{J} Y_i \log \hat{Y}_{t_i} + \lambda_1 \sum_{m=1}^{3}\sum_{n=1}^{N}\Big(1 - \frac{\sum\limits_{t=1}^{T_m}\alpha_{tnm}}{T_m}\Big)^2 + \lambda_2 \sum_{m=1}^{3}\frac{1}{T_m}\sum_{t=1}^{T_m}\Big(\sum_{n=1}^{N}\alpha_{tnm}\Big)^2. \tag{11}$$

其中:T_m为第m层ASGC-SRU网络层的时间步长;

Y_i 为第 i 类的真实标签; 式 (1) 中第 2 项为对不同的节点给予同等的关注; 式 (1) 中第 3 项为对感兴趣节点的数量进行限制; α_{tnm} 为 ASGC-SRU 网络第 m 层中, t 时刻第 n 个关节的注意力权重分数; λ_1、λ_2 均为权重衰减系数.

2.3　注意力增强空域图丢弃

传统的 Dropout 在图卷积计算中节点进行随机丢弃, 一定程度上缓解了过拟合问题, 如图 5(a) 所示. 由于手势运动中关节联动性极强, 即使一个节点被丢弃, 仍然可从其邻居节点中获得关于该节点的信息, 从而导致过拟合. 受 Cheng 等[20] 所做的人体动作识别研究的启发, 本文在 ASGC-SRU 的图卷积计算中引入空域图丢弃 (spatial dropgraph, SD) 来缓解丢弃节点间存在强关联性的问题.



(a) Dropout　　(b) Spatial Dropgraph

图 5　节点丢弃方式示意图

空域图丢弃的主要思想为: 当丢弃 1 个节点时, 同时会丢弃其邻居节点集. SD 有 2 个主要参数 p 和 K, 其中: p 控制样本概率, K 控制待丢弃邻居集的大小. 在输入特征图上, 本文首先用概率为 p 的伯努利分布对根节点进行采样, 然后丢弃采样的根节点和距离根节点为 K 的邻居节点, 如图 5(b) 所示.

对于 Dropout 而言, 直接以丢弃概率 μ 采样将被丢弃的节点, 对于 SD, 丢弃节点取决于参数 p 和 K. 给定一个具有 n 个关节点和 e 条边的手骨骼图, 估计其随机采样节点的 k 阶邻域中的期望节点数为

$$N_k = \bar{d} \times (\bar{d}-1)^{k-1}. \tag{12}$$

其中: $\bar{d}=\dfrac{2e}{n}$ 为节点的平均度, $k \in \{1, 2, \ldots, K\}$. 在手部骨骼图中, 需要丢弃的采样节点扩充至 k 阶邻域后的平均扩充倍数为

$$\bar{s} = 1 + \sum_{k=1}^{K} N_k. \tag{13}$$

于是, 手骨骼空域图丢弃的根节点采样概率为

$$p = \frac{\mu}{\bar{s}}. \tag{14}$$

接下来, 本文引入注意力增强机制对更冗余的根节点赋予更高的采样概率, 从而对其进行舍弃, 设计一种注意力增强空域图丢弃 (ASD). 由 Zagoruyko 等[21] 的思路可知, 对于离散图像数据而言, 在其特征通道维度上取绝对值的平均值, 能够增强图像中重要部分的体现. 但是对于多帧连续的空域手部骨骼图数据, 特别是样本量不足时, 这种重要性增强的思路往往会表现为增强了更冗余的节点特征, 该部分节点特征在网络的拟合过程中易加重过拟合, 故可对该部分节点进行丢弃处理.

本文将 ASGC-SRU 门结构中空域图卷积后的手势特征 $X \in R^{n\times T\times C}$, 在通道维度 C 和时间维度 T 上计算其绝对值的平均值, 得到注意力增强特征图 $\beta \in R^{n\times T\times C}$. 将注意力增强特征图归一化为

$$\hat{\beta} = \frac{\beta \times \mathrm{num}_\beta}{\sum \beta}, \tag{15}$$

其中 num_β 为 β 中的元素数量.

对归一化后的注意力增强图以概率 p 采样, 在节点数为 n 的手部骨骼图中, 将节点 i 采样为 ASD 的丢弃区域根节点的概率 p_i 为

$$p_i = \hat{\beta} p = \hat{\beta}\frac{\mu}{\bar{s}}. \tag{16}$$

ASD 算法流程描述如下.

step 1: 将 ASGC-SRU 门结构中图卷积计算后的特征 $X \in R^{n\times T\times C}$ 在时间 T 和通道维度 C 上计算其绝对值的平均值, 得到注意力增强特征图 β;

step 2: 归一化后的特征图 $\hat{\beta}$ 与 p 相乘得到 ASD 中第 i 个根节点采样概率 p_i;

step 3: 随机对 n 个样本的根节点进行采样, 每个采样根节点均服从概率为 p_i 的伯努利分布;

step 4: 结合邻接矩阵 A, 计算采样的根节点以及距其最大步长为 K 的节点的掩膜矩阵 M, 矩阵中所有采样节点值为 0;

step 5: 将手势特征 X 与矩阵 M 相乘, 进行掩膜计算, 丢弃采样节点;

step 6: 标准化特征 X 后从门结构中输出.

3　实验结果与比较

3.1　数据集与评价指标

SHREC'17 数据集[4] 是由英特尔 Realsense 相机采集的动态手势数据集. 包含由 28 名参与者分别以单指、全掌两种方式执行 1 ~ 10 次的 14 个手势序列, 共产生 2 800 个序列, 包含深度图视频帧序列以及 2D、3D 空间中 22 个手关节的坐标, 所提出方法基于 3D 骨骼坐标进行动态手势识别. 14 个类别的手势分别为抓取、轻敲、扩张、捏、顺逆时针旋转、上下左右滑动、X 型、十字型、V 型滑动、摆动. 每个手势既可用单指完成, 也可用全指完成, 这意味着手势可根据是否区分手指数量而分为 14 个类别或 28 个类别, 增

加了复杂程度和识别难度.

DHG 14/28 数据集[22] 包含由20名参与者以单指、全掌两种方式执行5次动作的14个手势的序列,类别同样可分为14类和28类,但该数据集提供了每个手势动作的关键帧区间.

对于SHREC'17数据集,使用与文献[4]相同的数据划分策略,将数据划分为1 980个训练数据和840个测试数据,并评估14类和28类手势的识别准确性. 在DHG 14/28数据集上,使用19:1的交叉验证策略进行评估[22].

3.2 实验细节

实验在一块NVIDIA GTX 1 080 GPU上进行,使用Pytorch 1.7实现模型搭建. 因数据集SHREC'17与DHG 14/28手势骨架序列长度不等,本实验使用线性插值法将骨架序列采样为统一长度T,均设为64,训练的批次大小分别设置为16和8. 网络中第1层全连接层通道数设置为128. 在4层ASGC-SRU中,隐藏层通道数设置为256,优化器使用Adam,初始学习率均为0.001,每20次迭代学习率减半,损失函数中λ_1、λ_2分别设为0.01和0.001. ASGC-SRU层间Dropout值设定为0.4,其中单元内部空域图卷积算子部分使用ASD取代Dropout,ASD参数在第3.3.2节中的实验设定.

3.3 实验结果与分析

3.3.1 注意力引导空域图卷积SRU有效性

本节在数据集SHREC'17上进行消融实验,验证SRU的快速性以及空域图卷积、注意力引导模块的有效性. 本文保持图1中数据输入后的运动特征融合增强不变,将之后的4层ASGC-SRU层分别替换为LSTM、SGC-LSTM、SRU、SGC-SRU,层数、参数均相同. 评估各方法在SHREC'17数据集中的识别准确率以及对测试集的840个动态骨架手势数据进行10次测试的平均识别时间,如表1所示.

表1 注意力引导空域图卷积SRU有效性

方法	14类/%	28类/%	t/s
LSTM	92.14	72.38	0.81
SGC-LSTM	93.57	88.70	1.90
SRU	93.10	71.79	0.64
SGC-SRU	94.05	88.45	1.10
ASGC-SRU	95.60	90.71	1.15

由表1可见,在14类的手势中SRU与LSTM精度相近. 在两者的基础上引入空域图卷积(SGC),由于14类手势不区分单指与全掌手势,手空间结构特征不需要过多区分,14类手势相较于LSTM、SRU提升了1.43%、0.95%,而在28类精细手势中引入空域图卷积后准确率有16.32%、16.66%的大幅提升,验证了引入空域图卷积显著增强对精细手势的识别能力. 在28类手势中,SGC-SRU性能比SGC-LSTM略低0.25%,在测试时间上,虽然手势特征融合增强、空域图卷积均会增加测试时间,但是SGC-SRU的测试效率提升为SGC-LSTM的1.72倍. 由此可见,SRU相较于LSTM,在几乎不损失手势识别精度的情况下大大提升了计算效率.

在此基础上引入指关节注意力引导模块,使得重要的手指关节具有更高的关注度,ASGC-SRU在14类和28类手势中准确率提升至95.60%、90.71%. 且注意力引导模块中的轻量注意力网络并不会对手势识别时间带来明显增加,但是对14类和28类手势的测试准确率提高了1.55%和2.26%.

由此可验证,ASGC-SRU在SRU的基础上引入空域图卷积、注意力引导模块,大大提升了网络对复杂手势的识别性能,同时保持较高的识别效率.

3.3.2 注意力增强空域图丢弃机制有效性

本节对比Dropout与注意力增强空域图丢弃(ASD)在SHREC'17上的表现. 设置$K=1$,即与根节点间步长为1的作为其邻接节点. 根据22个骨骼节点数和连接结构,由式(12)和(13)计算得到待丢弃采样根节点的平均度为2.27,节点采样平均扩充倍数为3.27. 对ASGC-SRU分别引入Dropout和ASD,节点扩充前的节点采样概率μ为两者共有参数,故在不同的μ值下进行对比,以选取具有最佳效果的μ值. 28类相比14类,总数据量不变,各类别数据量减半,更易导致过拟合,因此在28类手势的实验效果更明显. 本文对于不同的μ值,分别对两种方法进行3次实验并取其中最好的结果.

图6为28类手势中不同的采样概率下Dropout

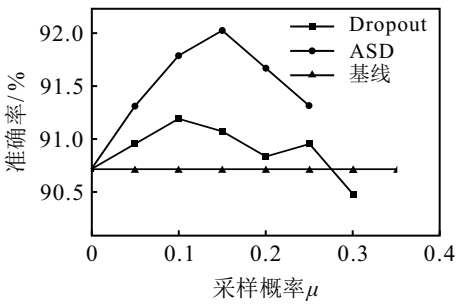


图6 不同节点采样概率μ下的对比

表2 注意力增强空域图丢弃机制有效性

方法	14类/%	28类/%
ASGC-SRU	95.6	90.7
ASGC-SRU + Dropout	95.7	91.2
ASGC-SRU + ASD	96.0	92.0

与ASD的对比. 实验在μ取0.05～0.2间具有有效的正则化效果,取0.15时具有最佳效果. 表2为14类和28类的结果,可见ASD对于更复杂的28类手势效果更显著.

3.3.3　方法比较与分析

SHREC'17和DHG 14/28数据集中同时提供了深度图像数据和骨骼坐标数据,本文仅选择基于3D骨骼坐标数据的方法进行对比,以保证对比方法使用原始数据的一致性.

表3为所提出方法与8种流行方法的对比结果. 所提出方法ASGC-SRU＋ASD在SHREC'17数据集上,对14类手势的识别准确率为96.0%,对更复杂的28类手势的识别准确率为92.0%.

表 3　在SHREC'17数据集上方法对比

方法	14类/%	28类/%
Smedt等[22]	88.2	81.9
MFA-Net [7]	91.3	86.6
ST-GCN[8]	91.6	86.9
STA-Res-TCN[13]	93.6	90.7
HG-GCN[15]	92.8	88.3
AGC-LSTM[9]	94.2	90.0
HPEV＋HMM＋FRPV[17]	94.9	92.3
Guo等[23]	94.8	92.9
ASGC-SRU＋ASD(本文)	96.0	92.0

图7和图8为14类手势和28类手势预测结果的混淆矩阵. 在SHREC'17的14类手势中有12个手势实现了高于90%的识别准确率,在28类手势中有20个手势实现了高于85%的识别准确率.

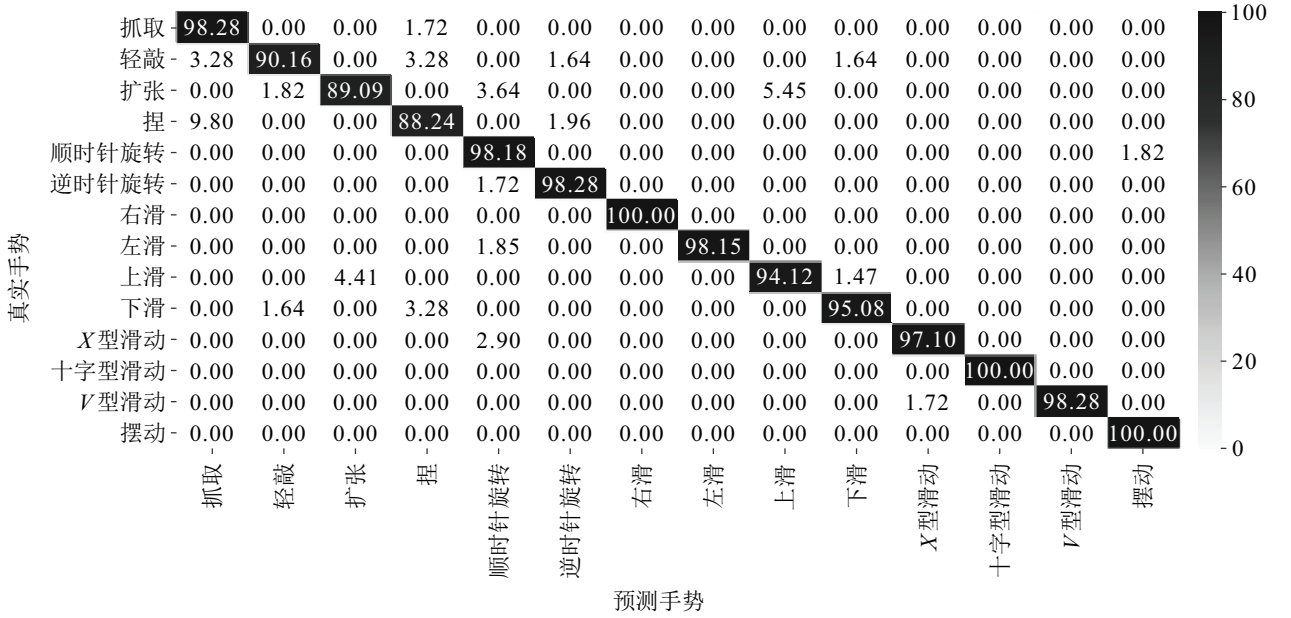


图 7　ASGC-SRU＋ASD在SHREC'17中14类混淆矩阵

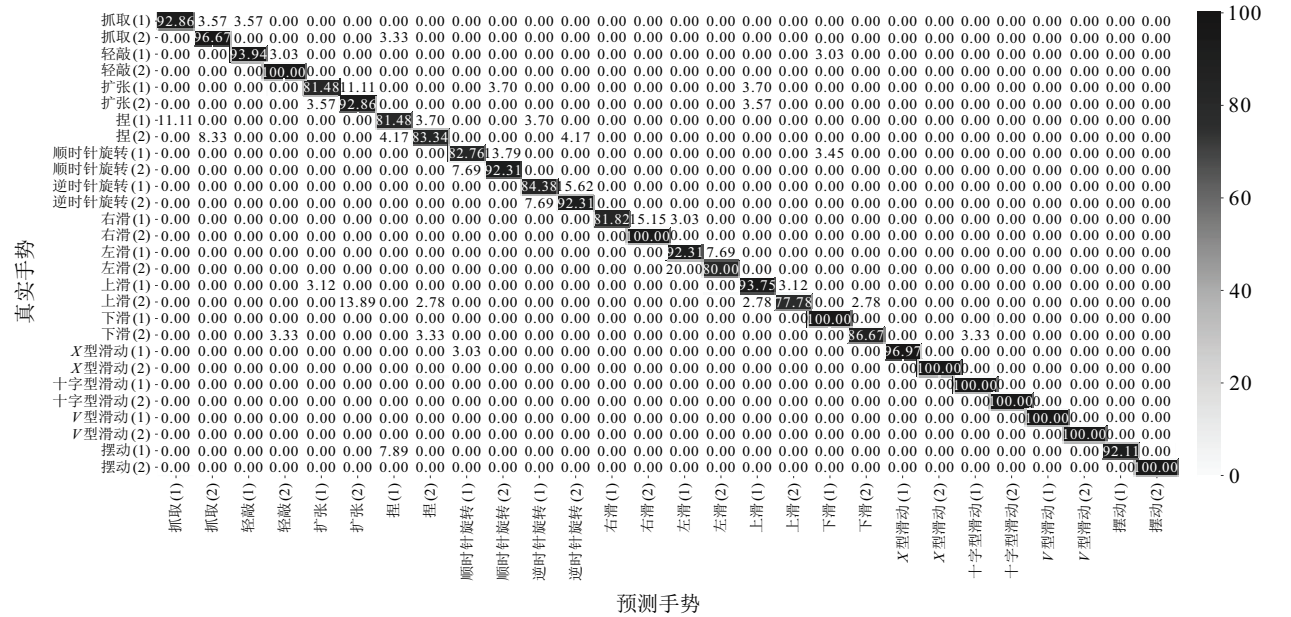


图 8　ASGC-SRU＋ASD在SHREC'17中28类混淆矩阵

在DHG 14/28数据集上进一步验证方法的有效性,与8种采用相同验证策略的方法进行对比,结果如表4所示. 所提出方法对14类手势的识别准确率为91.8%,对28类手势的识别准确率为88.3%.

表 4　在DHG 14/28数据集上方法对比

方法	14类/%	28类/%
SoCJ + HoHD + HoWR[4]	83.1	80.0
MFA-Net[7]	85.8	81.0
CNN + LSTM[12]	85.6	81.1
ST-GCN[8]	84.8	81.6
STA-Res-TCN[13]	89.2	85.0
HG-GCN[15]	89.2	85.3
AGC-LSTM[9]	88.7	86.1
HPEV + HMM + FRPV[17]	92.5	88.9
ASGC-SRU + ASD(本文)	91.8	88.3

图9和图10为DHG 14/28中14类和28类手势的混淆矩阵. 该数据集上所提出方法在14类手势中有10个手势实现了高于90.0%的准确率,在28类手势中有22个手势实现了高于85%的准确率.

此外,将所提出方法与主流方法在动态手势识别中的识别计算速率进行对比. 为保证实验配置的一致性,所提出方法与其他对比方法均在相同硬件算力的情况下进行对比. 因为方法之间输入待预测的手势数据采样帧数量存在一定差异,所以将每秒被识别计算的手骨架数作为速率评价指标,即每秒被识别计算的手骨架数(skeleton/s)等于每秒识别的手势数(gesture/s)乘以采样帧数(T).

图11为手势识别方法的计算速率对比结果. SoCJ + HoHD + HoWR[4]、Smedt等[22]的方法属于人工特征提取方法,实时性较差,故不进行速率比较.

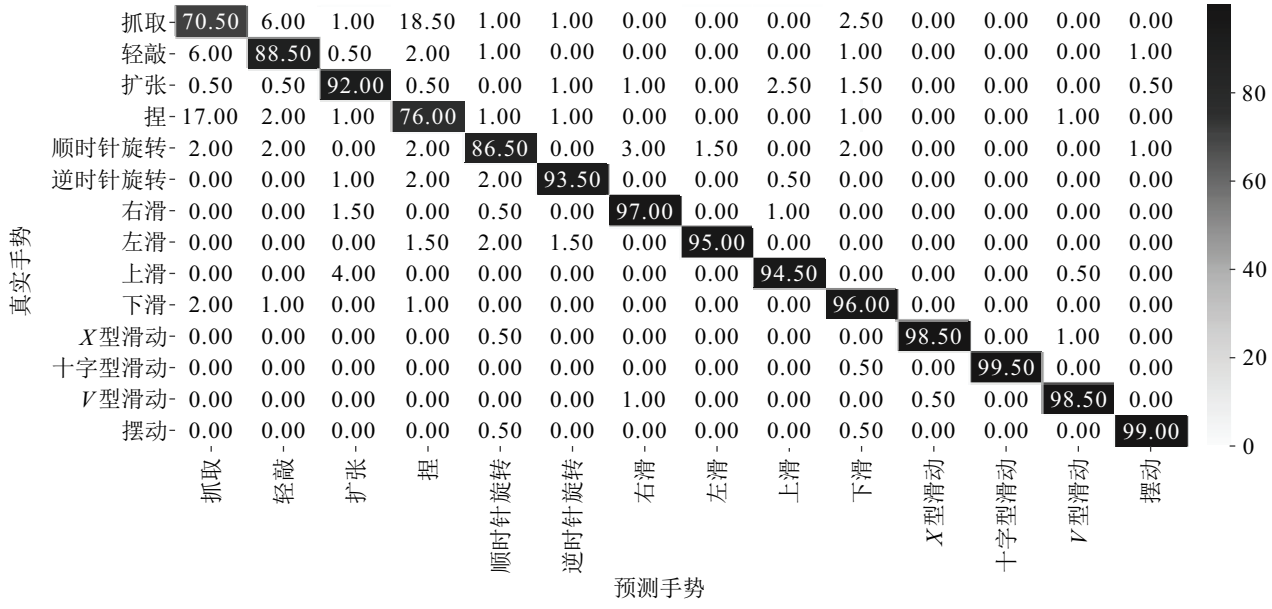


图 9　ASGC-SRU + ASD在DHG 14/28中14类混淆矩阵

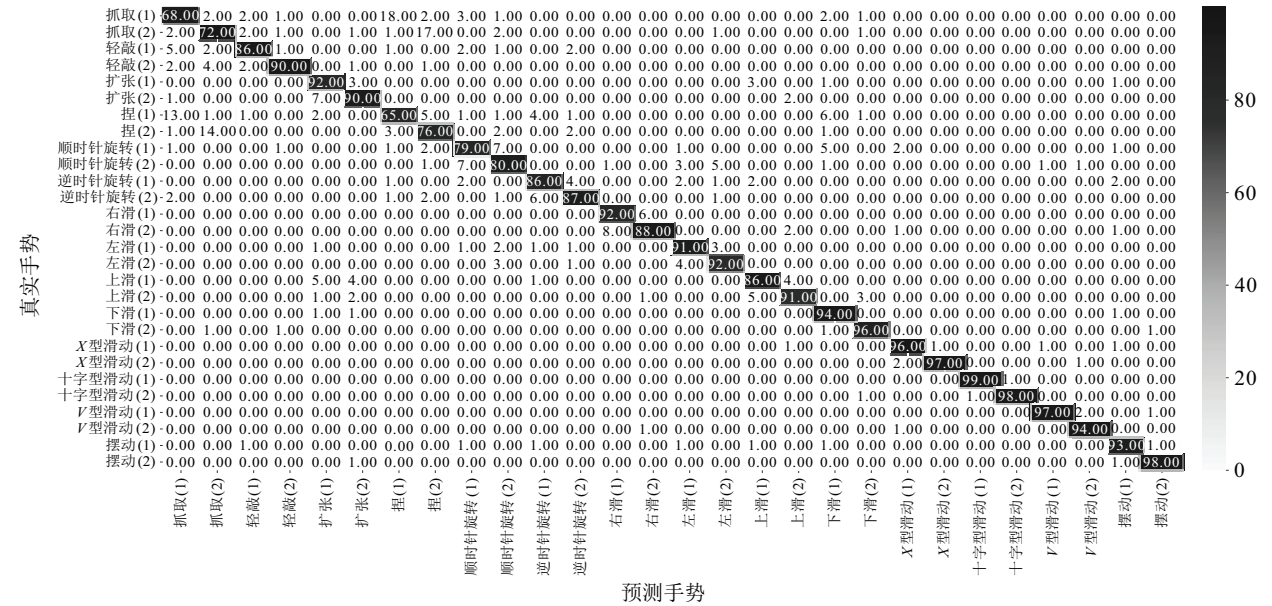


图 10　ASGC-SRU + ASD在DHG 14/28中28类混淆矩阵

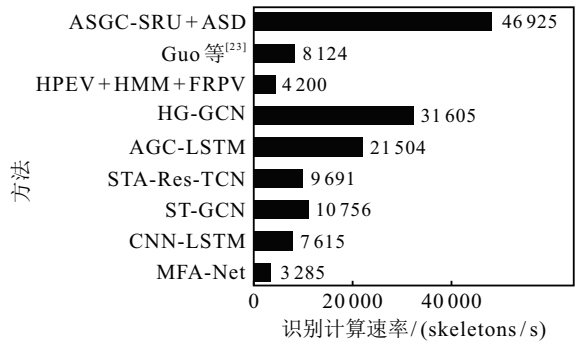


图 11　手势识别计算速率对比

由图11可见,所提出方法的识别计算速率达到了46 925 skeleton/s,显著高于其他方法. 其中与HG-GCN、AGC-LSTM、STA-Res-TCN、ST-GCN、CNN-LSTM、MFA-Net方法相比,不仅在识别准确率上高于这些方法,且识别计算效率远超STA-Res-TCN、ST-GCN、CNN-LSTM、MFA-Net,并达到了AGC-LSTM的2.18倍和HG-GCN的1.48倍.

由表3和表4可见,所提出方法在SHREC'17数据集的14类手势识别准确率最高,甚至高于HPEV+HMM+FRPV、Guo等[23]的方法1.1%、1.2%. Guo等[23]使用了一种边缘图卷积网络,HPEV+HMM+FRPV方法使用的是一种双流卷积网络,对提取的特征进行联合评估,这两种方法具有更大的参数量,降低了识别计算的快速性. 所提出方法在SHREC'17数据集的28类手势识别准确率仅低于这两种方法0.3%、0.9%. 在DHG 14/28上的14类和28类手势识别准确率分别仅低于HPEV+HMM+FRPV方法0.7%、0.6%.

通过同时对比识别准确率和计算速率,所提出方法在几乎不损失准确率的同时,识别计算速率能够显著高于其他方法,甚至达到Guo等[23]方法、HPEV+HMM+FRPV方法的5.8倍、11.2倍速率. 故可验证所提出方法在保证识别准确率具有一定优势的情况下,可显著地缩减识别计算时间,从而更直观地提升实际体验,且能够完全满足视频实时分析的需求(每帧30个手骨架).

4　结　论

本文提出一种基于注意力引导空域图卷积SRU(ASGC-SRU)的方法,用于基于手部骨骼的动态手势识别. 动态手势数据以图拓扑结构输入模型中进行拟合,在SRU的门结构中通过空域图卷积算子充分利用手势动作的空间结构信息,并配合多层高速并行的SRU高效地建模手势的复杂时空特征. 注意力引导模块成功地将指关节间的重要性程度加以区分,从而对基于不同手指的复杂手势具有更强的识别能力,并引入注意力增强空域图丢弃(ASD),减小随机节点丢弃带来的弊端,缓解了图卷积存在的过拟合问题. 在SHREC'17和DHG 14/28数据集上进行评估,验证了所提出方法在识别精度与计算效率间取得了很好的折衷. 下一步工作将着手RGB相机中通过手部姿态估计算法产生的三维骨骼图动态手势进行实时识别,以实现更低硬件成本、高精度的人机手势交互控制.

参考文献(References)

[1] Cihan Camgöz N, Koller O, Hadfield S, et al. Sign language transformers: Joint end-to-end sign language recognition and translation[C]. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos, 2018: 7784-7793.

[2] Chen T Z, Xu L T, Xu X S, et al. GestOnHMD: Enabling gesture-based interaction on low-cost VR head-mounted display[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(5): 2597-2607.

[3] He Z F, Zhang R F, Liu Z, et al. A user-defined gesture set for natural interaction in a smart kitchen environment[C]. The 13th International Symposium on Computational Intelligence and Design. Hangzhou, 2020: 122-125.

[4] de Smedt Q, Wannous H, Vandeborre J P. Skeleton-based dynamic hand gesture recognition[C]. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos, 2016: 1206-1214.

[5] de Smedt Q, Wannous H, Vandeborre J P. Heterogeneous hand gesture recognition using 3D dynamic skeletal data[J]. Computer Vision and Image Understanding, 2019, 181: 60-72.

[6] Weng J, Liu M, Jiang X, et al. Deformable pose traversal convolution for 3D action and gesture recognition[C]. Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 142-157.

[7] Chen X H, Wang G J, Guo H K, et al. MFA-Net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data[J]. Sensors, 2019,19(2): 239.

[8] 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. San Francisco, 2018: 7444-7452.

[9] Si C Y, Chen W T, Wang W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos, 2019: 1227-1236.

[10] She Q S, Mu G Y, Gan H T, et al. Spatio-temporal SRU with global context-aware attention for 3D human action

recognition[J]. Multimedia Tools and Applications, 2020, 79(17/18): 12349-12371.

[11] 赵俊男, 佘青山, 穆高原, 等. 基于MobileNetV 3与ST-SRU的危险驾驶姿态识别[J]. 控制与决策, 2022, 37(5): 1320-1328.
(Zhao J N, She Q S, Mu G Y, et al. Dangerous driving pose recognition based on MobileNetV 3 and ST-SRU[J]. Control and Decision, 2022, 37(5): 1320-1328.)

[12] Núñez J C, Cabido R, Pantrigo J J, et al. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition[J]. Pattern Recognition, 2018, 76: 80-94.

[13] Hou J X, Wang G J, Chen X H, et al. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition[C]. Proceedings of European Conference on Computer Vision. Heidelberg, 2018: 273-286.

[14] Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, 2017:156-165.

[15] Li Y, He Z H, Ye X, et al. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition[J]. EURASIP Journal on Image and Video Processing, 2019, 78(2019): 1-7.

[16] 缪永伟, 李佳颖, 孙树森. 融合手势全局运动和手指局部运动的动态手势识别[J]. 计算机辅助设计与图形学学报, 2020, 32(9): 1492-1501.
(Miao Y W, Li J Y, Sun S S. Dynamic gesture recognition combining global gesture motion and local finger motion[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(9): 1492-1501.)

[17] Liu J B, Liu Y C, Wang Y, et al. Decoupled representation learning for skeleton-based gesture recognition[C]. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos, 2020: 5751-5760.

[18] Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing, 2020, 29: 9532-9545.

[19] Lei T, Zhang Y, Wang S I, et al. Simple recurrent units for highly parallelizable recurrence[C]. Processing of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 4470-4481.

[20] Cheng K, Zhang Y F, Cao C Q, et al. Decoupling GCN with dropgraph module for skeleton-based action recognition[C]. Proceedings of European Conference on Computer Vision. Heidelberg, 2020: 536-553.

[21] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[C]. Proceedings of the International Conference on Learning Representations. Paris, 2017: 1-13.

[22] Smedt Q D, Wannous H, Vandeborre J P, et al. 3D hand gesture recognition using a depth and skeletal dataset: SHREC'17 track[C]. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, 2016: 33-38.

[23] Guo F T, He Z X, Zhang S Y, et al. Normalized edge convolutional networks for skeleton-based hand gesture recognition[J]. Pattern Recognition, 2021, 118(6): 108044.

作者简介

陈炫琦 (1998–), 男, 硕士生, 从事手势识别、人体动作识别等研究, E-mail: xqchen0505@foxmail.com;

佘青山 (1980–), 男, 教授, 博士, 从事动作识别、主动康复机器人、脑机交互、机器学习等研究, E-mail: qsshe@hdu.edu.cn;

张波涛 (1982–), 男, 副教授, 博士, 从事智能机器人与智能控制、图像处理及其应用等研究, E-mail: billow@hdu.edu.cn;

马玉良 (1976–), 男, 教授, 博士, 从事机器学习、模式识别、图像处理、脑-机接口技术等研究, E-mail: mayuliang@hdu.edu.cn;

张建海 (1978–), 男, 教授, 博士, 从事人工智能、信号处理、脑-机接口等研究, E-mail: jhzhang@hdu.edu.cn.