高广荣, 李云峰. 基于 SAGC-GRU 的动态手势识别研究 [J]. 智能计算机与应用,2025,15(6):81-89. DOI:10.20169/j.issn.2095-2163.250612

基于 SAGC-GRU 的动态手势识别研究

高广荣, 李云峰
(河南科技大学 机电工程学院, 河南 洛阳 471003)

摘 要: 传统动态手势识别方法的特征提取通常难以充分利用手部关节之间的运动学依赖关系。针对现有的动态手势识别方法提取时空特征尺度单一、有效特征的提取能力不足问题,提出了一种基于自注意力的图卷积门控循环网络(SAGC-GRU)的动态手势识别方法。首先利用图卷积网络(GCN)和自注意力机制,同时关注手势的全局和局部信息,对不同空间位置上的特征进行加权,浅层特征和深层特征相结合,其次将增强的空间特征输入到门控循环单元(GRU)中,对时间维度上的信息进行建模,从而获得序列特征表示。在数据集 SHREC-2017 中对 14 类手势进行了测试实验,并与其它方法进行对比分析,实验结果表明识别准确率达 97.3%,在计算精度上具有一定的优势。

关键词: 动态手势识别; 时空特征; 门控循环单元; 图卷积网络; 注意力机制
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2025)06-0081-09

Research on dynamic gesture recognition based on SAGC-GRU

GAO Guangrong, LI Yunfeng
(School of Mechatronics Engineering, Henan University of Science and Technology, Luoyang 471003, Henan, China)

Abstract: Traditional dynamic gesture recognition methods with feature extraction are usually difficult to fully utilize the kinematic dependencies between hand joints. To address the problems that existing dynamic gesture recognition methods extract spatio-temporal features at a single scale and the extraction capability of effective features is insufficient, the paper proposes a Self-Attention Guided Graph Convolutional Gated Recurrent Unit Network (SAGC-GRU) method for dynamic gesture recognition. Firstly, the Graph Convolutional Network (GCN) and the self-attention mechanism are used to pay attention to the global and local information of gestures at the same time, and the features in different spatial positions are weighted, and the shallow features and deep features are combined. Then the enhanced spatial features are input into the Gated Recurrent Unit (GRU) to model the information in the time dimension to obtain the sequence feature representation. Test experiments on 14 types of gestures are conducted in the dataset SHREC-2017 and compared and analyzed with other methods, and the experimental results show that the recognition accuracy reaches 97.3%, which is an advantage in terms of computational accuracy.

Key words: dynamic gesture recognition; spatio-temporal feature; gated recurrent unit; graph convolutional network; self-attention mechanism

0 引 言

当今,计算机技术逐渐应用到社会各个领域中。而在各项计算机技术中,动态手势识别作为一种直观、便捷的人机交互方式,为人们提供了与计算机系统无缝沟通的便利,在高速发展的信息化社会中发挥着重要作用。动态手势识别在手语认知领域[1]、智慧交通[2]、医疗健康[3]、智能工业[4-6]等领域中有着潜在的应用价值。根据手势的特点和应用场景的差异,手势识别可以分为静态手势识别[7]和动态手势识别[8]两种类型。其中,静态手势识别通过对手的姿态或手部形态的图像特征进行提取和分类。动态手势识别根据手势形状和空间位置随时间的变化来识别手势所表达的信息。相比之下,动态手势识别需要考虑手势在时间序列中的空间运动信息,通常需要对手势进行连续跟踪和识别,技术研发更具

基金项目: 河南省重大科技专项(221100220100)。
作者简介: 高广荣(1996—),女,硕士研究生,CCF 会员,主要研究方向:图像处理,计算机视觉。
通信作者: 李云峰(1973—),男,博士,教授,硕士生导师,主要研究方向:图像处理,计算机视觉。Email:liyunfeng379@126.com。
收稿日期: 2023-11-06

有挑战性,实现起来更加复杂。

目前,动态手势识别正成为国内、外相关技术领域关注的热点。科研人员尝试利用手部外观特征[9]或手势运动信息[10]进行动态手势识别。其中,基于外观特征的手势识别方法通常是通过分析手势图像[11]来实现的。然而,基于外观特征的方法容易受到光线、阴影等环境因素的影响,同时也无法适用于一些轮廓不够明显或不太规则的手势,例如抓握、捏合等动作。基于运动信息的手势识别利用手势的骨架模型[12]的时空演变进行动态手势识别。基于手部骨骼的方法是利用手部关键节点,通过其相对的位置信息来描述手势。这样可以直观地表示手势的形状和动态变化,避免了对整张图像进行复杂处理的需求,使得运行速度更快,且不会受到背景、光照变化等干扰,具有更好的鲁棒性。

基于骨架模型的手势识别的传统方法通常是由人工构建手势骨架序列中的特征,并将特征运用于动态手势的分类识别。Smedt 等学者[13]提出了一种基于 3D 骨架数据的异构手势识别方法,从不同视角获取的 3D 骨架数据中提取特征并融合在一起。随着深度学习技术的发展,基于骨架的手势识别方法受到了广泛关注。Lai 等学者[14]提出了 CNN 与 RNN 串联融合并结合骨架信息,以提取手势的时空特征。这些方法将手势骨架数据建模为一系列向量,通过 CNN 或 RNN 进行处理以直接学习手部的时空特征,然而这种方法无法充分利用手部关节之间的运动学依赖关系。

Wang 等学者[15]提出构建基于 ST-GCN 的时空图卷积网络模型的思路,对实时帧图像进行分割和特征提取,利用组合网络获得稳定准确的手骨关键点。Zhao 等学者[16]提出了基于骨架的动态手势识别局部时空同步网络(LSTSN),通过注意力机制和自适应阈值方法,实现了有效的时空信息捕捉和复杂局部相关性建模。袁冠等学者[17]提出了基于时空图神经网络的手势识别算法,使用传感器的空间位置信息和图神经网络表征手势数据的空间关联性,同时采用 GRU 解决手势的时序性和长距离依赖问题。陈炫琦等学者[18]提出了 ASGC-SRU 网络,利用空域图卷积和 SRU 的门结构,并引入指关节注意力和注意力增强空域图丢弃正则化方法,提高了手势识别的准确性。Song 等学者[19]提出了一种基于多流改进时空图卷积网络的动态手势识别方法,采用自适应空间图卷积和多种膨胀率的扩展时间图卷积,结合时空和通道注意力层进行手势识别。Li 等学者[20]提出了一个端到端的手势图卷积网络,其中卷积操作仅在连接的骨架关节上进行,同时通过扩展坐标维数来增加训练数据集中的语义特征。

利用手势骨架关键节点的位置信息构成图结构数据,而图卷积网络(Graph Convolutional Network, GCN)针对图结构数据提取手势骨架关键节点的空间特征信息,通过图卷积训练得到的特征能够更准确地匹配手势识别任务的要求。为了研究动态手势识别中的时空特征,结合门控单元(Gated Recurrent Unit,GRU)在处理特征时间维度方面具有独特优势,本文提出了基于自注意力的图卷积门控循环网络(Self-Attention Guided graph Convolutional Gated Recurrent Unit network,SAGC-GRU)。图卷积网络(GCN)同时关注手势骨架关键节点的全局和局部信息,对不同位置节点的特征进行加权,浅层特征和深层特征相结合。将这些加强的空间特征输入到门控单元(GRU)中,以实现对时间维度上的信息建模,从而获得序列特征表示。但并不是所有节点的特征对于当前时刻的手势分类任务都是重要的,因此引入自注意力机制,使模型能够自适应地选择最相关的节点特征进行信息传递。

1 SAGC-GRU 网络

1.1 网络模型结构

本文所提出的 SAGC-GRU 网络模型结构如图 1 所示,主要包含手部图构造、手势时序特征提取、手势识别三个部分。首先,使用多帧动态手势骨骼序列作为网络模型的输入进行手部图结构构建。然后,手部图通过多层 GCN 融合自注意力机制增强数据节点权重,将骨骼节点的三维坐标数据向高维特征空间进行特征映射,并对相邻帧间的差异特征进行拼接,以丰富数据的特征信息,每层 GCN 采用平均池化,同时嵌入空域图卷积算子,可以更好地捕捉手骨骼数据的空间上的特征。为了解决手势识别过程中时序性和长依赖的问题,网络经过特征变换,按照时间顺序将同一个骨骼节点上不同采样空间特征转化为时间序列,并借助门控循环单元(GRU)提取手势数据的时序特征。最后,通过全连接层(FC)将注意力权重分数加权后的隐藏状态与网络单元末端隐藏状态输出进行聚合,最终通过 Softmax 函数实现手势预测。

1.2 手部图结构建立

构建动态手势骨骼序列的拓扑结构,以便更好地理解手部骨架的关联关系。手部图结构构建如图

2 所示，根据手部骨架的生理结构构成无向图，手指关节的运动通常由靠近掌心一侧的节点来控制远离掌心的关节。因此，将手部骨骼空间图中最靠近掌心的节点定义为根节点 r。给定时间帧 t 下，手部关节点 n 表示为 $u_{tn}=(x_{tn},y_{tn},z_{tn})$，手部骨架的拓扑结构可表示为一个无向空间图 $G_t=(U_t,E_t)$，集合 U_t 包括所有的关节点，表示为 $U_t=\{u_{t1},u_{t2},u_{t3},\cdots,u_{tn}\}$。其中，$n$ 表示节点的数量。边集合 E_t 由手部骨骼的关节点间的无向连接组成，即 $E_t=\{(u_{ti},u_{tj})\mid u_{ti},u_{tj}\in U_t\}$。为定义每个节点的邻接节点集合，引入了采样距离的概念，节点 u_{ti} 外接节点集合被定义为 $N_d(u_{ti})=\{u_{tj}\mid d(u_{ti},u_{tj})\leqslant 1\}$，$d(u_{ti},u_{tj})$ 表示从节点 u_{ti} 到节点 u_{tj} 最短距离，采样距离默认小于 1。关节点集距离只描述了每对关节点之间的距离，为了进一步研究手势骨骼的方向信息，采用关节点方向锥（Joint Direction Cone，JDC）。从关节点集中选择一个点为起始点、编号为 k 来测算起始点与其他点的偏移，如果手腕关节点为起始点，那么在 t 骨架关节点方向锥特征的计算公式可以表示为：

$$JDC^t=(\overrightarrow{j_k^t\,j_1^t},\overrightarrow{j_k^t\,j_2^t},\cdots,\overrightarrow{j_k^t\,j_N^t}) \tag{1}$$

其中，$\overrightarrow{j_i^t\,j_j^t}=(x_{ti}-x_{tj},y_{ti}-y_{tj},z_{ti}-z_{tj})$。

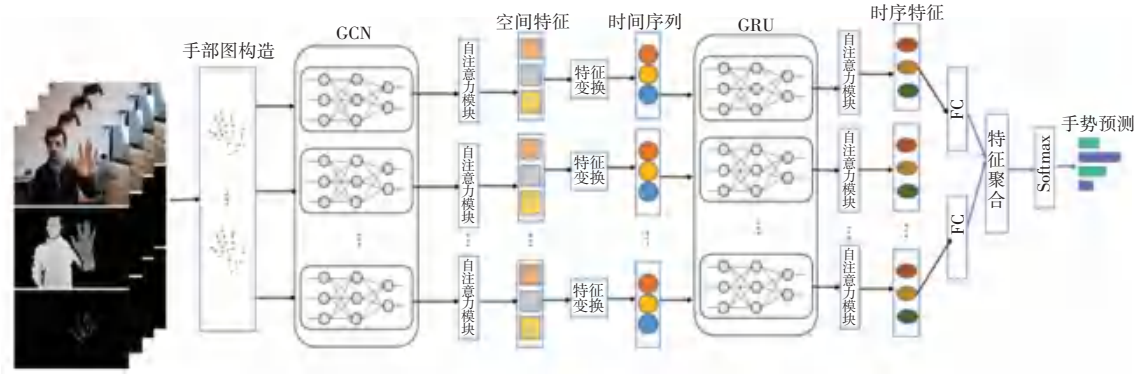


图 1　模型总体框架

Fig. 1　General framework of the model

根据以上定义，分析手势动作的运动特征、关节点之间的相对位置、方向以及手部骨架的整体结构，可以利用无向空间图 G_t 来描述动态手势骨架的拓扑结构，并通过节点之间的边界连接关系来捕捉手部骨骼之间的关联。

图 2　手部图结构构建

Fig. 2　Structural construction diagrams of hands

1.3　手部空域图卷积

为了解决非欧式空间数据的特征提取问题，采用图卷积神经网络来建模手部骨骼数据，以挖掘骨骼间的空间关系。空域图卷积主要是对节点的邻居节点集合中的采样函数和权重函数进行拓展。

采用基于标签的策略来定义适用于空间域的权重函数，以适应图数据结构的多样性。首先将邻接节点集合划分为 K 个子集，接着为每个子集分配一个数字标签。定义标签函数为：

$$G_S(N_d(u_{ti}))=j,\ j=0,1,\cdots,k-1 \tag{2}$$

其中，$N_d(u_{ti})$ 表示节点 u_{ti} 外接节点集合。则权重函数定义为：

$$W(u_{ti},u_{tj})=W'(G_{ti}(u_{tj})) \tag{3}$$

其中，$G_{ti}(u_{tj})$ 表示标签函数。则空域图卷积可定义为：

$$Y_{\text{space}}=\sum_{u_{tj}\in Nd(u_{ti})}\frac{1}{z_{ti}(u_{ti})}X(d(u_{ti},u_{tj}))\cdot W(u_{ti},u_{tj}) \tag{4}$$

其中，$d(u_{ti},u_{tj})$ 表示节点 u_{ti} 与节点 u_{tj} 间距离；$z_{ti}(u_{ti})=|\{u_{tk}\mid G_{ti}(u_{tk})=G_{ti}(u_{tj})\}|$ 表示归一化项，确保平衡不同子集对输出的贡献。综合上述公式，在 t 时刻 u_{ti} 卷积输出表示为：

$$Y_{\text{space}}=\sum_{u_{tj}\in Nd(u_{ti})}\frac{1}{z_{ti}(u_{ti})}X(u_{tj})\cdot W(G(u_{tj})) \tag{5}$$

1.4　门控循环单元

动态手势数据具有时序性和长依赖性的特性，这意味着手势数据会随着时间的推移而发生变化，不仅会受到当前时间的影响，而且还受到之前某一时间点输入数据的影响。为了更好地提取手势数据中的时序信息，使用门控循环单元（GRU）对动态手

势在时间流角度进行特征分析,GCN-GRU 网络将同一关节点在不同的采样位置上的空间特征转化为一系列时间序列特征。手势时序特征如图 3 所示,这样一来在保留空间特征的同时,就能更加充分地捕捉时间上的演变趋势。

图 3 手势时序特征

Fig. 3 Gesture timing features

GRU 有 2 个门来控制信息的输入与输出,即 1 个重置门(Reset Gate)和 1 个更新门(Update Gate)。重置门和更新门具有长期记忆能力,可以保存在长时间序列中的信息,并不会随着时间的推移而丢失或清除。运行中,重置门和更新门相互配合,最终筛选出门控循环单元的最终输出信息。GRU 的结构图 4 所示。

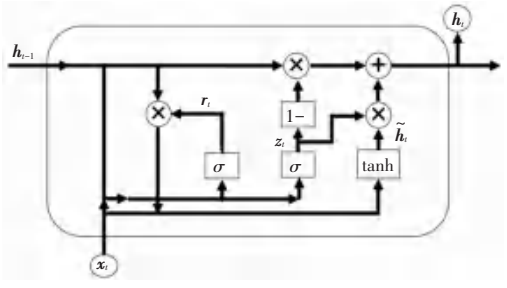


图 4 GRU 内部结构

Fig. 4 Internal structure of GRU

首先,将时序特征 $\boldsymbol{x}_t$ 前一时刻的隐藏状态 $\boldsymbol{h}_{t-1}$ 进行线性变换后相加在一起,形成 1 个更新门。隐状态 $\boldsymbol{h}_{t-1}$ 融合了上一个时间步的相关信息。该门计算得到更新状态 $\boldsymbol{z}_t$,用来控制前一时刻和当前时间之间的信息传递。这种机制是循环神经网络中最关键的组成部分,能够在序列数据中捕捉长期依赖关系。其计算公式为:

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{6}$$

其中,σ 表示 Sigmoid 函数。通过 Sigmoid 函数将输入数据范围限制在 0 到 1 之间的范围内,生成门控信号;$\boldsymbol{W}_z$ 表示权重矩阵参数;$\boldsymbol{b}_z$ 表示偏置。然后设置重置门控制前一时刻的隐藏中各维度的信息有哪些需要被遗忘,即哪些手势信息应该被传递到当前时刻的计算,需用到的公式为:

$$\boldsymbol{r}_t = (\boldsymbol{w}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{7}$$

接着计算新记忆状态 $\boldsymbol{h}_t'$;新的记忆内容使用重置门储存过去相关的信息。可由如下公式计算求出:

$$\boldsymbol{h}_t' = \tanh(\boldsymbol{w}_h' \cdot [\boldsymbol{r}_t \cdot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h') \tag{8}$$

新记忆状态 $\boldsymbol{h}_t'$ 融合了当前输入的 $\boldsymbol{x}_t$ 数据。有针对性地更新了当前时刻的隐藏状态,即记忆了当前时刻的状态信息。最后,网络需要计算 $\boldsymbol{h}_t$,该向量保存当前单元的状态信息并传递给下一个节点的隐藏状态,可用下式进行计算:

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \cdot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \cdot \boldsymbol{h}_t' \tag{9}$$

其中,$\boldsymbol{z}_t$ 表示更新门的激活结果,是以门控的形式控制信息的流入。$\boldsymbol{z}_t$ 与 $\boldsymbol{h}_t'$ 的矩阵乘积表示了前一刻时间需要保留至最终记忆的信息,将其与当前时刻保留至最终记忆的信息相加,就得到了门控循环单元最终的输出内容。

1.5 自注意力机制

在基于手骨骼的动态手势识别中,对于一些较为精细的手势,只有部分关节会起主导作用。为了更好地捕捉这些关键关节的特征,引入基于时空位置编码的自注意力模块,设计结构如图 5 所示。通过自注意力机制对时间序列进行建模,计算序列中每个元素和其他元素之间的相似度得分,并根据相似度分配不同的注意力权重。

时空位置编码是由空间位置编码 PE_s(Spatial Positional Encoding)和时间位置编码 PE_t(Temporal Positional Encoding)两部分组成的,都使用了正弦和余弦函数来设置其值。具体来说,空间位置编码是将一个 N 维的空间位置向量映射到一个 N 维的位置编码向量,该向量的每个维度都由正弦和余弦函数计算而来,这样可以为不同的位置赋予不同的编码向量,并保留位置之间的相对关系。时间位置编码则类似于空间位置编码,将时间序列中的每个时间步都映射为一个向量,也是由正弦和余弦函数计算得到。研究得到的数学公式为:

$$\begin{cases} PE(pos, 2i) = \sin\left(\dfrac{pos}{10\,000^{2i/d_{\text{model}}}}\right) \\ PE(pos, 2i+1) = \cos\left(\dfrac{pos}{10\,000^{2i/d_{\text{model}}}}\right) \end{cases} \tag{10}$$

其中,pos 表示每个关节点的位置;i 表示位置编码的维度;d_{model} 表示模型的维度。空间位置编码中的 pos 由连续的 N 维向量组成,时间位置编码中的 pos 由 $N \times M$ 维向量组成,为手势序列中的每个关节点的位置赋予了一个对应的值。这些空间位置编码的向量将会与原始节点特征向量相加,经过自

注意力层进行归一化处理。处理后的节点特征再与时间位置编码合并，用于标记这些节点特征在手势序列时间轴上的位置。更新后的节点特征还将经过一次多头自注意力层处理，以便模型能够更好地学习手势关节点的时空特征，计算公式如下：

$$H_{\text{out}} = ATT(PE_s + ATT(H_{\text{in}} + PE_t)) \quad (11)$$

其中，H_{out} 表示最终的输出特征；ATT 表示多头自注意力层；H_{in} 表示初始的输入特征。

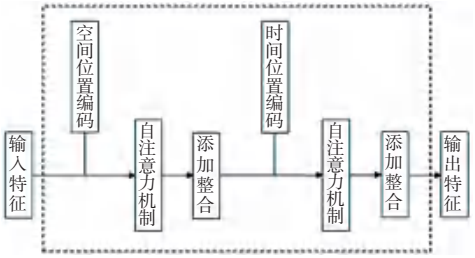


图 5　自注意力机制模块

Fig. 5　Self-attention mechanism module

2　实验

为了验证时空动态手势识别的准确率和性能，采用动态手势经典数据集 SHREC-2017[21-22] 进行训练。首先对提出的方法进行消融实验来验证方法的有效性，然后将方法与其他经典动态手势识别基线方法进行对比试验，来验证所提出方法的优势。

2.1　数据集选择

采用公开的动态手势数据集 SHREC-2017 进行实验。该数据集共包含 2 800 个视频段，涵盖 14 种手势，并通过手部运动的复杂性和细节程度来定义、共分为粗粒度和细粒度两类手势。其中，粗粒度手势是指由相对简单的动作组成的手势，如上下、左右、前后等基本动作，这些手势通常由较少的关键帧（Key Frame）组成。细粒度手势则是由更复杂的动作组成的手势，包括旋转、弯曲、挤压等动作。2 种不同的手型如图 6 所示，每种手势又可使用单指或全掌手势方式演示。完整的手部骨架如图 7 所示。数据集包含深度图像和骨骼两种数据模态，骨骼数据的关节点坐标分为 2D 深度图像中 22 个手关节的二维坐标和 3D 世界坐标系下 22 个手关节的三维坐标。本实验使用的是骨骼数据的三维坐标信息。数据集提供了官方划分、即按照 7∶3 的比例进行训练集与测试集的划分，其中训练集有 1 960 段数据，测试集有 840 段数据。

（a）一根手指的手势　　（b）整只手的手势

图 6　2 种不同的手型

Fig. 6　Two different hand patterns

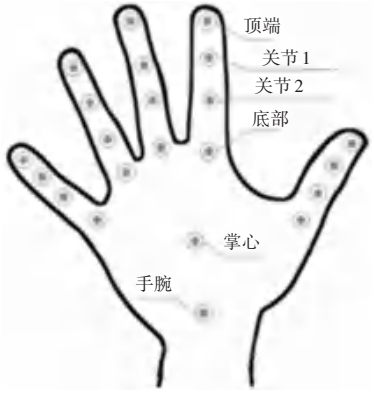


图 7　完整的手部骨架

Fig. 7　Complete hand skeletons

动态手势是序列动作、即连续的。为了保证动态手势视频序列关键帧是有效的，减少冗余帧，实验采用数据集以 30 帧/s 为间隔提取得到手势动作关键帧，得到手势样本的长度为 20~30 帧。运动骨骼图像如图 8 所示。以捏动作为例，提取关键样本帧过后的手势骨骼关键帧运动轨迹图。对于三维手部骨架序列，每一帧骨架数据包含 22 个三维坐标信息，将每一帧骨架样本节点相连作为一个图结构，作为网络输入。

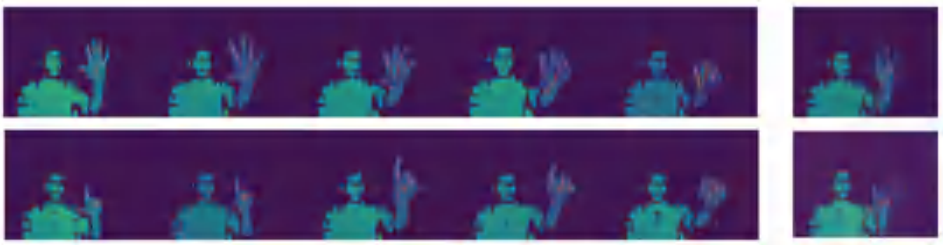


图 8　运动骨骼图像

Fig. 8　Motion skeletal imaging

2.2 实验环境与实验设置

实验基于 Pytorch 实现,在微软 Windows10 操作系统中进行训练,训练过程中使用 NVIDIA GeForce RTX 2080 Ti 显卡进行加速,具体实验环境见表 1。

表 1 实验环境信息

Table 1 Experimental environment information

环境名称	环境配置及版本
CPU	Intel(R)Xeon(R)CPUE5-2678v3@ 2.50 GHz
内存	64 G
显卡	NVIDIA GeForce RTX 3050 Ti
操作系统	Windows10 专业版
网络框架	Pytorch1.7 深度学习框架
Python 版本	3.8.13
CUDA 版本	10.1

实验使用 GCN 网络模型来提取 3D 手势骨骼空间特征,其中 GCN 层数和 GRU 层数在后续参数分析中进行确定。在训练过程中设置批处理大小为 35,使用 Adam 作为训练优化器,总迭代次数(epoch)设置为 200,batch size 设置为 32,初始学习率设置为 0.01,分别按照 0.01 速率在迭代次数(epoch)为 10、40、70、100 进行衰减。本实验的评价指标使用准确率,用于判断网络对手势的识别程度的优劣。准确率(*Accuracy*)是指对于给定的测试手势数据集,识别模型正确分类器的样本数占总样本数的比值。具体计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

3 结果分析

3.1 网络结构

图卷积神经网络通过聚合相互连接的手势关键点来提取不同粒度的空间特征,来表征手势的不同层次的语义信息。浅层特征包含更多的原始信息,但存在语义歧义的问题;深层特征具有较高的语义性,可以有效地表征原始数据,但可能会丢失原始数据的特性。因此通过自注意力机制的融合,可以有效地表征原始数据,并避免深层特征丢失原始数据的特性的问题。同时,选择不同的图卷积神经网络的层数也会影响手势识别的性能。实验结果如图 9 所示。通过多次实验发现,在增加网络层数时,手势识别准确率也随之增加,但当层数增加到 3 时,准确率逐渐趋于平稳,此后再增加可能出现下降趋势。因此,采用了深度为 3 的图卷积神经网络来提取手势的空间特征。

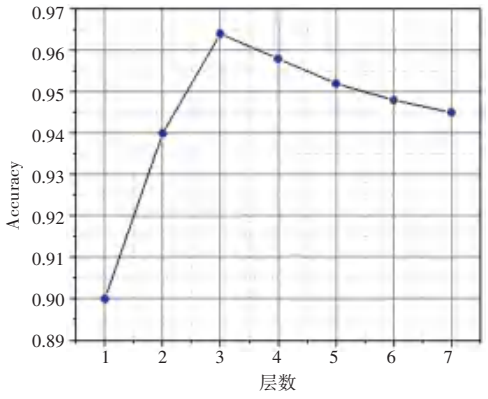


图 9 图卷积层数影响

Fig. 9 Effects of the number of graph convolution layers

在手势识别任务中使用了不同深度的 GRU 网络来提取手势序列的时间特征,并将这些特征输入到全连接层进行分类。通过增加 GRU 网络的层数发现,在网络层数从 1 增加到 3 的过程中,手势识别性能呈现逐步提升的趋势。这是因为增加 GRU 网络的层数可以提高模型的表征能力,从而更好地捕捉手势序列的时序信息。然而,当进一步增加网络层数时,模型性能开始出现下降趋势。这是因为在网络变得更深时,会增加训练难度,导致网络出现梯度消失或梯度爆炸等问题,从而影响模型的性能表现。实验结果见表 2。在手势识别任务中,适当增加 GRU 网络的深度可以提升模型性能,但需要注意网络过深可能会带来训练问题,因此需要根据实际情况来确定网络的深度。

表 2 GRU 网络层数叠加实验结果

Table 2 Experimental results on the superposition of GRU network layers

方法	准确率/%
AGCN+GRU	96.1
AGCN + GRU+ GRU	96.6
AGCN +GRU+ GRU+ GRU	97.3

3.2 消融试验

为了验证所提出的融合注意力机制的有效性,对 3 种不同的网络模型进行了消融实验,分别是:仅使用图卷积的 GCN 模型、加入门控单元的 GCN+GRU 模型,以及融合自注意力机制的图卷积与门控单元模块的 AGCN-GRU 模型。实验结果见表 3。分析表 3 实验结果可知,GCN 网络的准确率为 94.6%,加入门控单元的 GCN+GRU 准确率提升至 96.1%,提升了 1.5%,而加入融合注意力机制的 AGCN-GRU 网络准确率提升最为明显,达到了 97.3%,提升了 2.8%。这表明,在手势识别任务中,加入融合注意力机制的图

卷积与门控单元模块的 AGCN-GRU 网络相比于其他 2 种网络结构具有更好的性能和有效性。

表 3　自注意力机制的有效性实验结果

Table 3　Experimental results on the effectiveness of the self-attention mechanism

方法	准确率/%
GCN	94.6
GCN+GRU	96.1
AGCN+GRU	**97.3**

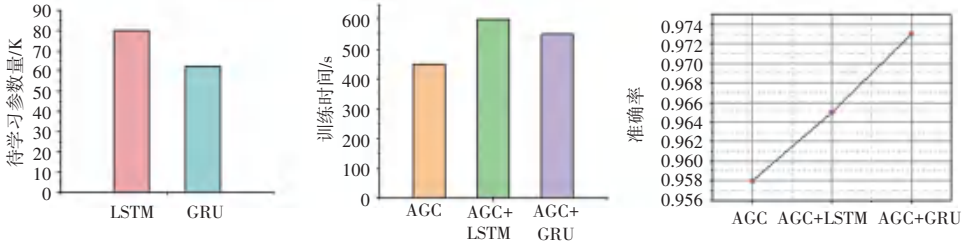
实验还对 GRU 的有效性进行分析，并设计了 3 种不同的模型：只使用图卷积神经网络融合注意力机制的、加入时间门控 LSTM 的模型、以及加入时间门控 GRU 的模型。实验结果见表 4。由表 4 可知，加入时间门控机制的模型的性能要优于只使用图卷积神经网络的模型。具体而言，只使用图卷积神经网络融合注意力机制的模型的准确率为 95.1%。而加入时间门控 LSTM 之后，模型的准确率提升了 1.47%，达到 96.5%。更为显著的是，加入时间门控 GRU 之后，模型的准确率增加了 2.31%，达到了 97.3%，更直观的准确率提升幅度如图 10(c)所示。这就表明，相较于 LSTM，GRU 更为有效，因为 GRU 中的门控机制更加灵活，减轻了梯度消失或梯度爆炸的问题，能够更好地处理时间序列数据。

表 4　GRU 有效性实验结果

Table 4　Results of GRU validity experiments

方法	准确率/%
AGCN	95.1
AGCN+LSTM	96.5
AGCN+GRU	**97.3**

实验还比较了使用 GRU 和 LSTM 进行手势识别时，所需要的待学习参数数量的差异。实验结果见图 10(a)，可知使用 LSTM 的模型需要 80 k 的待学习参数，而使用 GRU 的模型只需要 62 k 的待学习参数。因此，在手势识别中使用 GRU 可以有效地减少模型中待学习的参数数量，从而提高模型的泛化能力和训练速度。这是因为 GRU 相比于 LSTM 具有待学习参数更少、训练时间更短的特点。两者模型训练时间如图 10(b)所示，可知 GRU 的模型训练时间低于 LSTM。GRU 单元中的参数相比于 LSTM 更少，在长时序的手势数据上训练得到的模型泛化能力更优，在手势识别任务中使用 GRU 网络可以提高模型的效率和性能。



(a) SHREC 2017 数据集上的待学习参数量　(b) SHREC 2017 数据集上的训练时间　(c) SHREC 2017 数据集上的准确率

图 10　GRU 有效性

Fig. 10　GRU effectiveness

3.3　对比试验

混淆矩阵如图 11 所示，展示了模型在 SHREC-2017 数据集 14 个手势类别上的预测结果混淆矩阵。分析混淆矩阵可以看到模型在 SHREC-17 数据集的 14 个手势类别中，有 12 个手势实现了高于 90%的识别准确率。这表明该模型在识别不同的手势方面具有良好的性能。

为了进一步验证该模型的识别有效性，实验将其与另外 6 种同样使用手势骨骼识别方法进行对比。在确保对比方法使用的原始数据一致性的基础上，得到实验结果见表 5。AGCN-GRU 模型的准确度为 97.3%，准确率均优于其他基线方法。准确率对比如图 12 所示，可看到准确率柱状图的结果对比更加明显。

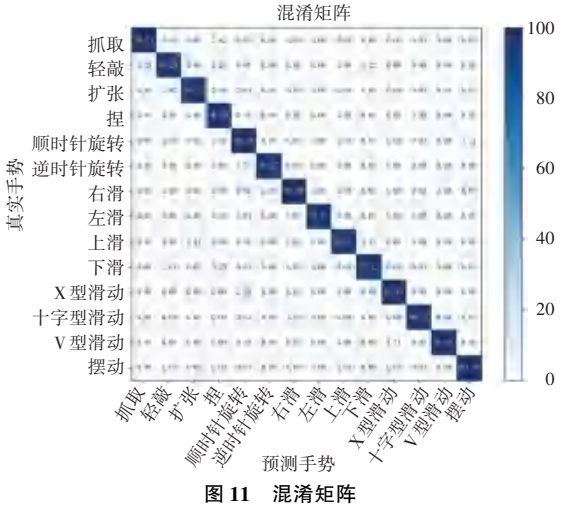


图 11　混淆矩阵

Fig. 11　Confusion matrix

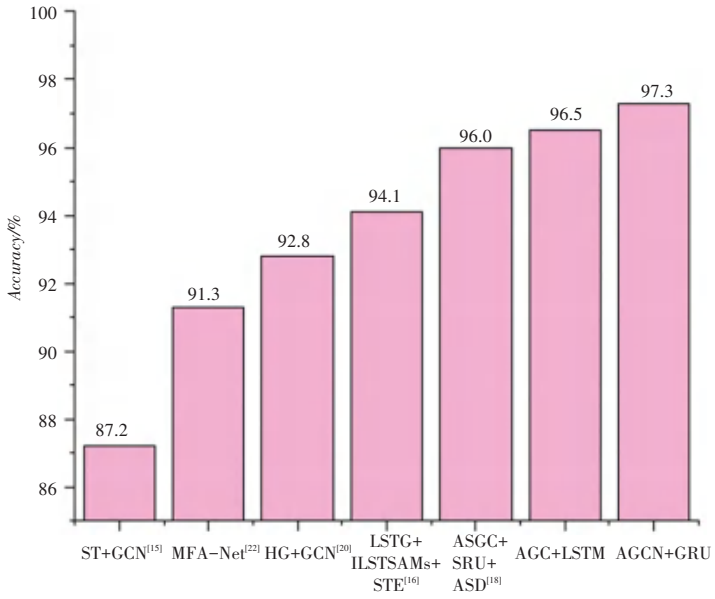


图 12 准确率对比图

Fig. 12 Comparison results of accuracy

表 5 准确率对比实验结果

Table 5 Accuracy comparison experimental results

方法	准确率/%
ST+GCN[15]	91.6
MFA-Net[22]	91.3
HG+GCN[20]	92.8
LSTG+ILSTSAMs+STE[16]	94.1
ASGC+SRU+ASD[18]	96.0
AGC+LSTM	94.2
AGCN+GRU	**97.3**

4 结束语

本文提出了一种 AGCN-GRU 的网络模型,用于识别动态手势。首先进行手势骨骼的图结构建立,作为 AGCN-GRU 模型的输入。为了解决手势时序性问题,将经过图卷积处理得到的空间特征作为门控单元 GRU 的输入,利用节点位置信息进行编码,同时引入自注意力机制增强节点权重,以增强网络对手势识别的能力。实验表明,该模型在动态手势识别中相对于基线方法准确度有所提升。由于目前在手势时序特征研究时只考虑了单方向的时序特征,在未来的研究上可以进一步丰富数据,并在时序性方面探索双向时间关系。

参考文献

[1] SUNDAR B, BAGYAMMA T. American sign language recognition for alphabets using MediaPipe and LSTM[J]. Procedia Computer Science, 2022, 215: 642-651.

[2] 李泰国, 张英志, 张天策, 等. 基于改进 YOLOv5s 算法的列车驾驶员手势识别[J]. 铁道学报, 2023, 45(1): 75-83.

[3] LI Runqing, ZHENG Diwei, HAN Ziyi, et al. mHealth: A smart phone-controlled, wearable platform for tumour treatment[J]. Materials Today, 2020, 40: 91-100.

[4] GAO Xiaojie, JIN Yueming, DOU Qi, et al. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search[C]// Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway, NJ:IEEE,2020: 8440-8446.

[5] QI Wen, LIU Xiaorui, ZHANG Longbin, et al. Adaptive sensor fusion labeling framework for hand pose recognition in robot teleoperation[J]. Assembly Automation, 2021, 41(3): 393-400.

[6] INKULU A K, BAHUBALENDRUNI M V A, DARA A, et al. Challenges and opportunities in human robot collaboration context of industry 4.0 - a state of the artreview[J]. Industrial Robot: The International Journal of Robotics Research and Application, 2022, 49(2): 226-239.

[7] ADITHYA V, RAJESH R. A deep convolutional neural network approach for static hand gesturerecognition[J]. Procedia Computer Science, 2020, 171: 2353-2361.

[8] GAO Q, CHEN Y, JU Z, et al. Dynamic hand gesture recognition based on 3D hand pose estimation for human-robot interaction[J]. IEEE Sensors Journal, 2022, 22(18): 17421-17430.

[9] AL-HAMMADI M, MUHAMMAD G, ABDUL W, et al. Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation[J]. IEEE Access, 2020, 8: 192527-192542.

[10] JIANG D, LI G, SUN Y, et al. Gesture recognition based on skeletonization algorithm and CNN with ASL database[J]. Multimedia Tools and Applications, 2019, 78(21): 29953-29970.

[11] SANTOS C, SAMATELO J, VASSALLO R. Dynamic gesture

recognition by using CNNs and star RGB：A temporal information condensation [J]. Neurocomputing, 2020, 400：238-254.
[12] KAPUSCINSKI T, MIS M. Differential pseudo - image for skeleton-based dynamic gesture recognition[C]// Proceedings of 2022 IEEE International Conference on Image Processing (ICIP). Piscataway,NJ:IEEE, 2022：4203-4207.
[13]SMEDT D 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.
[14]LAI K, YANUSHKEVICH S N. CNN+RNN depth and skeleton based dynamic hand gesture recognition[C]// Proceedings of the 24th International Conference on Pattern Recognition. Beijing：Chinese Academy of Sciences, 2018：3451-3456.
[15]WANG Jingyao, YU Naigong, FIRDAOUS E. Gesture recognition matching based on dynamic skeleton [C]// Proceedings of the 33rd Chinese Control and Decision Conference (CCDC). Piscataway,NJ:IEEE,2021：1680-1685.
[16]ZHAO Dongdong, YANG Qinglian, ZHOU Xingwen, et al. Temporal synchronous network to dynamic gesture recognition [J]. IEEE Transactions on Computational Social Systems, 2023, 10(5)：2226-2233.
[17]袁冠，邴睿，刘肖，等. 基于时空图神经网络的手势识别[J]. 电子学报，2022，50(4)：921-931.
[18]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.
[19]SONG J, KONG K, KANG S. Dynamic hand gesture recognition using improved spatio - temporal graph convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9)：6227-6239.
[20]LI Yong, HE Zhang, YE Xiang, et al. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition [J]. EURASIP Journal on Image and Video Processing, 2019, 2019：78.
[21]SMEDT D Q, WANNOUS H, VANDEBORRE J P. Skeleton-based dynamic hand gesture rcognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ：IEEE, 2016：1-9.
[22]CHEN Xinghao, WANG Guijin, GUO Hengkai, et al. Mfa-net：Motion feature augmented network for dynamic hand gesture recognition from skeletal data[J]. Sensors (Basel), 2019, 19(2)：239.