融合时空注意力的改进ST-GCN人体动作识别方法研究

雷建云[ab],梁钧[ab],夏梦[ac*],张慧丽[ac],田祚汉[ab]

(中南民族大学 a.计算机科学学院; b.湖北省制造企业智能管理工程技术研究中心;c.农业区块链与智能管理湖北省工程研究中心,武汉 430074)

摘 要 针对现有的人体骨架动作识别算法不能充分发掘运动的时空特征问题,提出了一种基于融合时空注意力的改进图卷积网络模型.该模型包含空间注意力机制和时间注意力机制,利用时空注意力机制从时间和空间两个维度分别提取动作的全局时空特征.将这二者融合到统一的时空图卷积网络(ST-GCN)框架中,实现了端到端的训练.在Kinetics和NTU RGB+D两个公开数据集的对比实验证明:改进模型在NTU-RGB+D数据集上的CS标准下取得了82.37%的Top-1精度,在CV标准下取得89.84%的Top-1精度,相比原来的ST-GCN算法,分别提升0.87%的Top-1精度和1.54%的Top-5精度.在Kinetics数据集上,改进模型取得了31.78%的精度,与ST-GCN相比提高了1.08%.由此验证了改进方法的有效性.

关键词 图卷积网络;骨架数据;动作识别;时空注意力

中图分类号 TP391.41 **文献标志码** A **文章编号** 1672-4321(2025)04-0526-10

doi:10.20056/j.cnki.ZNMDZK.20250412

Research on the improved ST-GCN method for human action recognition by integrating spatiotemporal attention

LEI Jianyun[ab],LIANG Jun[ab],XIA Meng[ac*],ZHANG Huili[ac],TIAN Zuohan[ab]

(South-Central Minzu University, a. College of Computer Science; b. Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprise; c. Hubei Provincial Engineering Research Center of Agricultural Blockchain and Intelligent Management, Wuhan 430074, China)

Abstract Aiming at the problem that existing human skeleton action recognition algorithms cannot fully explore the spatiotemporal features of motion, an improved graph convolutional network model based on fusion of spatiotemporal attention is proposed. This model includes spatial attention mechanism and temporal attention mechanism, utilizing spatiotemporal attention mechanism to extract global spatiotemporal features of actions from both temporal and spatial dimensions. Integrating these two into a unified spatiotemporal graph convolutional network (ST-GCN) framework enables end-to-end training. Comparative experiments on two publicly available datasets, Kinetics and NTU RGB+D, have shown that the improved model achieves a Top-1 accuracy of 82.37% under the CS standard on the NTU RGB+D dataset, and a Top-1 accuracy of 89.84% under the CV standard. Compared with the original ST-GCN algorithm, the improved model achieves a Top-1 accuracy of 0.87% and a Top-5 accuracy of 1.54%, respectively. On the Kinetics dataset, the improved model achieved an accuracy of 31.78%, which is 1.08% higher than ST-GCN. This validates the effectiveness of the improved method.

Keywords graph convolutional network; skeleton data; action recognition; temporal and spatial attention

收稿日期 2024-03-19 ***通信作者** 夏梦,研究方向:深度学习与图形识别,E-mail:xiameng@mail.scuec.edu.cn

作者简介 雷建云(1972-),男,教授,博士,研究方向:大数据与网络空间安全,E-mail:leijianyun@mail.scuec.edu.cn

基金项目 新疆维吾尔自治区区域协同创新专项资助项目(科技援疆计划)(2022E02035);湖北省技术创新计划重点研发专项资助项目(2023BAB087);武汉市重点研发计划资助项目(2023010402010614);武汉市知识创新专项曙光计划资助项目(2023010201020465)

近年来，计算机视觉作为人工智能领域的一个重要分支，也是最热门的研究领域之一，已经在许多领域得到广泛应用. 简单来说，计算机视觉，又或者说机器视觉，就是研究计算机如何模拟人类大脑去解释和处理使用各种各样的硬件设备，如摄像头等模拟人眼“看”到的目标图像. 因此衍生出计算机视觉领域中两个热点研究方向：物体识别和人体动作识别[1].

人体动作识别是指对视频或图像中的人体姿态进行分析与理解，提取其特征并通过算法完成动作分类的过程. 该技术在民生领域及军事、公共安全领域得到广泛应用，促进了各行业智能化产品的发展，加快了社会的进步. 人体动作识别既有长远的理论研究意义，又在人机交互、智慧看护、体育健身等民生领域，以及智能监控等公共安全领域逐渐普及使用，进一步给广大人民的日常生活带来诸多便利.

1　相关工作

随着科技的快速发展，各种高精度深度摄像设备层出不穷，再加上姿态估计算法[2]，二者配合便可较为容易得到骨架数据. 因此，基于骨架数据的人体动作识别方法也不断发展. 一开始是传统方法，后来因深度学习的兴起，研究主流逐渐往深度学习方向靠拢.

在人体动作识别研究早期，传统方法的特点是需要研究人员自行设计手工特征，再以此特征对人体进行建模. 例如，VEMULAPALLI 等[3]使用关节轨迹的协方差矩阵；FERNANDO 等[4]使用关节点的相对位置；LIU 等[5]则是平移或旋转人体之间的各部分.

而基于深度学习的方法主要使用卷积神经网络（Convolutional Neural Networks，CNN）和递归神经网络（Recurrent Neural Networks，RNN）两种广泛流行的模型以端到端的方式进行学习. 在基于CNN的方法中，KE 等[6]将关节点间距离映射到图像上，再输入CNN模型实现动作分类. LI 等[7]设计了一种骨架变换器模块实现自动重排、选择重要骨架关节. 而在基于RNN的方法中，DU 等[8]利用分层RNN组合、融合不同身体部位的特征，以实现最终的动作预测. SONG 等[9]建立具有长短期记忆的RNN基础模型，应用注意力机制分别对不同帧和关节点赋予不同程度的关注.

人体骨架是一种图结构数据，而图结构数据是一种典型的非欧式结构数据，其排列无序性、边具有额外信息等特点给传统卷积方式迁移到图结构数据上带来了挑战. 而基于图的模型[10]也因此应运而生. 当前图模型主要包括两种形式，一种是图神经网络（Graph Neural Network，GNN）. GNN 结合图与RNN，通过消息传递与多次迭代来捕获节点间语义关系和结构信息[11]. QI 等[12]将GNN应用于图像、视频检测和人机交互任务；LI 等[13]利用GNN建模物体之间的依赖关系等.

另一种是图卷积网络（Graph Convolutional Network，GCN）. GCN将CNN扩展到了图模型.KIPF 等[14]引入光谱GCN进行半监督分类；SIMONOVSKY 等[15]将GCN用于点云分类. YAN 等[16]首次将GCN用于人体动作识别，提出时空图卷积网络（Spatial-Temporal Graph Convolutional Networks，ST-GCN），开创了GCN在人体动作识别领域研究应用的先河，后续诸多研究都是以该网络为基础开展的.

本文主要工作如下：引入时空注意力机制来学习人体骨架的时空特征，并将其有机地融合到时空图卷积网络ST-GCN模型中，得到融合时空注意力机制的改进ST-GCN模型，并实现端到端训练. 实验结果表明，本文方法在动作识别任务中取得了较好的效果.

2　融合时空注意力的改进ST-GCN模型

2.1　时空图卷积ST-GCN

在ST-GCN中，通过姿态估计算法或kinectis相机捕获人体骨架图，并将其转化为关节点坐标信息，进而根据骨骼数据构建相应的连接关系. 将构建好的人体关节点图序列数据输入到时空图卷积网络ST-GCN中，进行空间和时间两个维度上的图卷积以提取更高级的特征图. 最后，通过全连接层、分类器等进行分类，输出最终的结果. 整体流程如图1所示.

分别在时间和空间两个不同的维度上，对骨架序列中的关节点进行建模[17]，如图2的时空图所示. 深灰色的点代表当前帧下的人体关节点，浅灰色的点则是代表相邻帧（即当前帧的前一帧或后一帧）下的人体关节点，蓝色的线条代表人体关节点之间的现实连接，如头部和脖子、左手和左肩膀、右手和

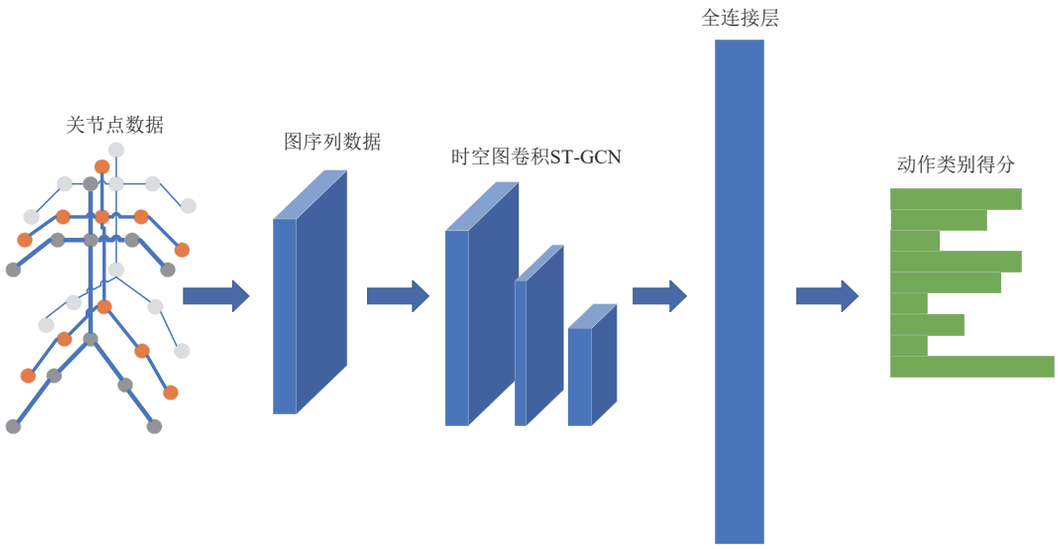

图 1 ST-GCN 识别流程

Fig. 1 ST-GCN identification process

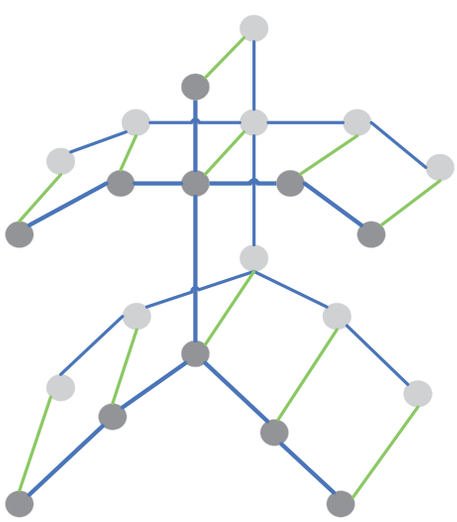

图 2 时空图

Fig. 2 Time-Space diagram

右肩膀等等. 绿色的线条代表相同人体关节点在相邻帧之间的时间连接,如右手节点在前一帧的位置、此时的位置、后一帧的位置的连接. 用数学语言来翻译,可以表述为输入一个人体关节点序列(N, T),其中N表示人体的N个关节点,T表示输入序列的长度. 以此为基础构建无向图$G=(V,E)$,V代表图节点集合,即$V=\{v_{ti}|t=1,2,\cdots,T,i=1,2,\cdots,N\}$. E代表边集合,由E_s和E_t两个部分组成,E_s表示在当前帧上人体关节点的现实连接,E_t表示不同帧下相同关节点的时间连接.

由上段中定义好的无向图G,接着定义在当前帧下(空间维度层面)图卷积运算. 对于第τ帧上的节点$V_{\tau i}$,它可以表示为:

$$f_{\text{out}}(v_{\tau i})=\sum_{v_{\tau j}\in s_i}\frac{1}{T_{ij}}f_{\text{in}}(v_{\tau i})w(l_i(v_{\tau i})), \tag{1}$$

其中:v代表图G的节点,f_{in}代表特征映射,s_i是目标节点$v_{\tau i}$卷积的采样区域,权重函数w用来提供权重向量,映射函数l为特征向量分配权重. s_i的大小不同,所包含的子集数也不同. 如图 3 所示,O 表示骨架的重心,红色区域内是s_i,s_i由 3 个子集组成:s_{i1}是目标节点本身(红色圆圈),s_{i2}是向心节点集(黄色圆圈),s_{i3}是离心节点集(蓝色圆圈). 每个子集有自己的标签和映射l_i.T_{ij}表示顶点$v_{\tau j}$所在s_i子集的基数. 对公式进行转换,得到图卷积在空间维度上实现的公式为:

$$f_{\text{out}}=\sum_{k}^{Kv}w_v(f_{\text{in}}(\tilde{\boldsymbol{A}}_k\odot\boldsymbol{M}_k)), \tag{2}$$

其中:K_v表示卷积核的大小,$\tilde{\boldsymbol{A}}$是邻接矩阵$\boldsymbol{A}$的归一

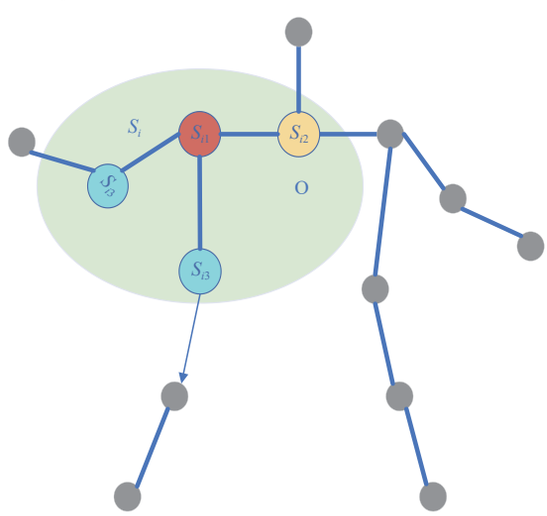

图 3 图卷积

Fig. 3 Graph convolutional

化形式，$\boldsymbol{M}$是一个可学习的权重矩阵，⊙符号表示点积.

接下来定义在相邻帧（即当前帧的前一帧或后一帧）上的图卷积. 对于每个顶点v_{ti}，以其为中心，向前一帧和向后一帧看，其对应的相同关节点有且仅有那两个，也就是说在整个人体关节点序列中，从时间层面来看，每个人体关节点都有两个固定的邻居节点. 因此，仅需要对模型输出的特征图进行二维卷积即可在时间维度上完成图卷积运算操作.

原始ST-GCN网络模型总体结构如图4所示. 整个网络包含了10层ST-GCN模块，除第一个ST-GCN模块外，后9个ST-GCN模块不仅包括图卷积与时间卷积模块，还包括了残差网络. 特征图经Polling层、FC层处理完成后，进入Softmax分类器，最终输出结果.

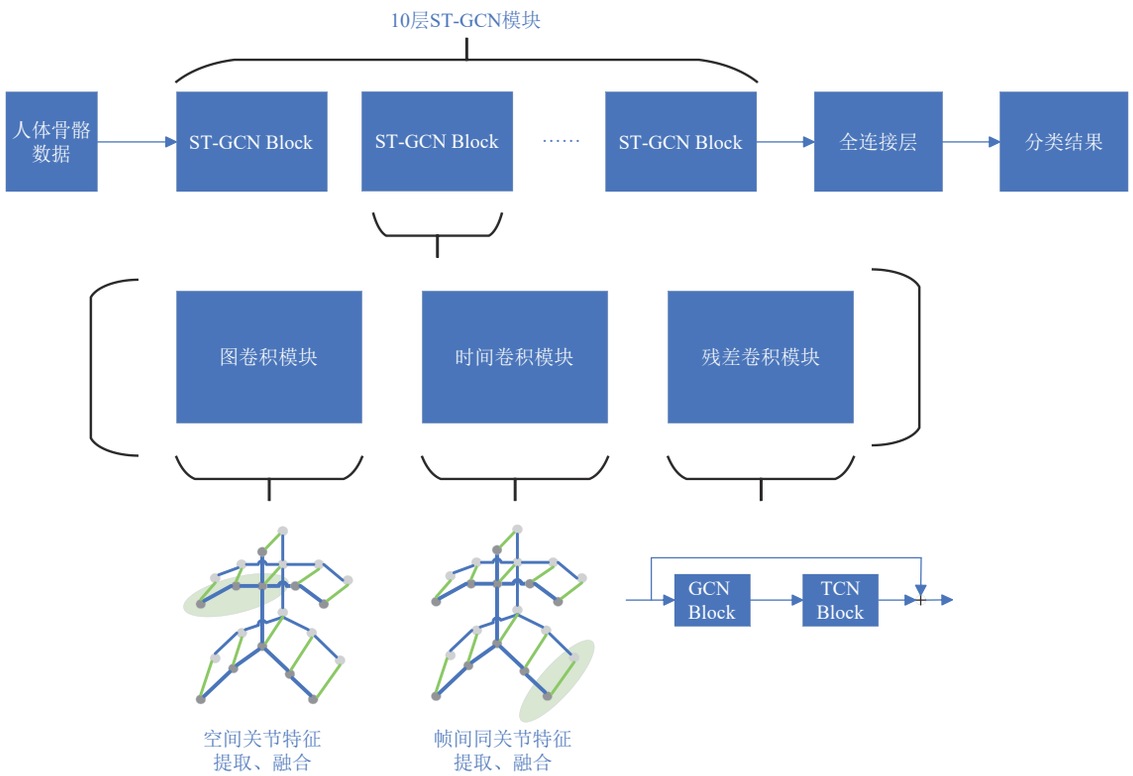

图4　ST-GCN总体结构

Fig. 4　Overall structure of ST-GCN

2.2　融合时空注意力的改进ST-GCN设计

2.2.1　融合时空注意力的ST-GCN结构

本文在原模型ST-GCN模块的基础上，添加时空注意力模块，从时间维度和空间维度两个提取人体关节点特征并进行融合，强化了特征图的全局特征信息. 改进后的ST-GCN总体结构如图5所示，红框部分为改进内容.

本文的时空注意力模块如图6所示，上路为时间注意力子模块：该子模块首先通过一个1×3的卷积层对输入进行处理，用于学习视频序列在时间维度上的权重分布. 然后经过Batch Normalization层，用于规范化输出的数据分布. 最后通过Sigmoid激活函数，将输出限制在0到1之间的范围内，用于表示每个时间步的权重. 下路是空间注意力子模块，功能和时间注意力模块类似，不同之处在于使用一个3×1的卷积层对输入进行处理，最后输出的是每个空间位置的权重. 将二者融合最终得到包含时间和空间信息的融合特征表示. 通过时空注意力模块的学习，特征图的全局特征信息得到了强化.

融合了时空注意力机制的改进后ST-GCN模型结构如图7所示. 该网络包含10个基本单元，起始通道数为3. 利用OpenPose算法提取好的人体关节点骨架信息在输入基本单元以前，会首先经过BN层，做归一化处理以增强数据规范性. 除第一个基本单元外，从第二个基本单元开始到最后一个基本单元，均使用残差机制进行链接. 可以根据输入的通道将基本单元分为3组，第1组即输出通道数为64，步长为1，如图7蓝色单元部分所示；第2组即输出通道数为128，步长为1，如图7绿色单元部分所示；第3组即输出通道数为256，步长为1，如图7灰色单元部分所示. 除个别单元外，如第4个和第6个基本单元数据输入输出通道数不改变，但步长会由1变

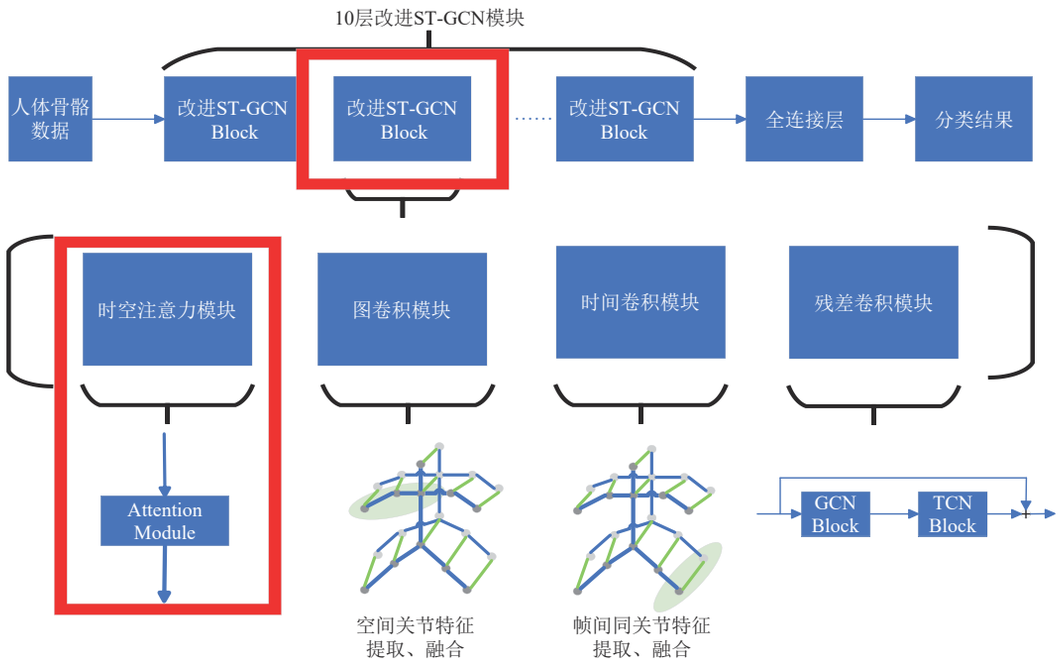

图5 改进ST-GCN总体结构

Fig. 5 Overall structure of improved ST-GCN

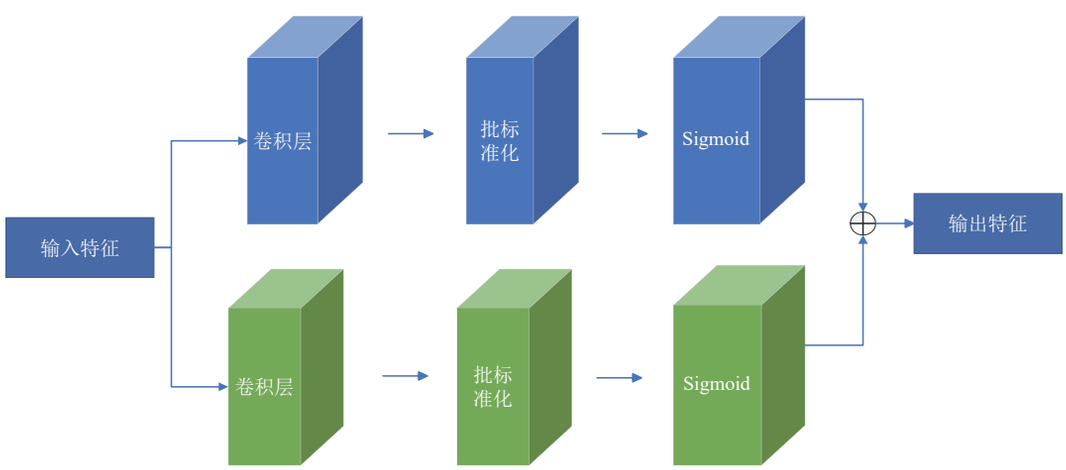

图6 时空注意力模块

Fig. 6 Spatiotemporal attention module

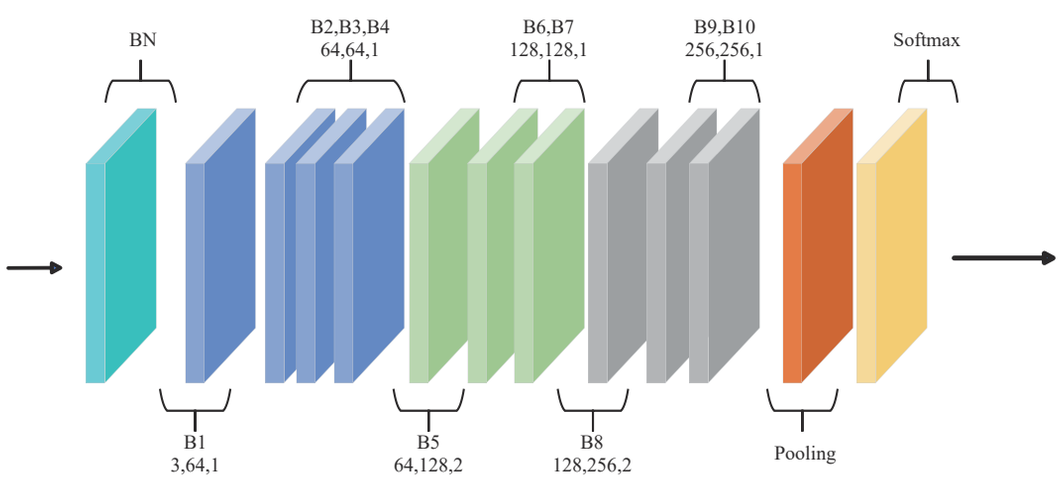

图7 改进ST-GCN模型层数结构

Fig. 7 Structure of improved ST-GCN layers

为2. 经过上述所有单元处理后，特征图经过pooling层、FC层，最后由Softmax分类器输出最终结果.

2.2.2　融合时空注意力的改进ST-GCN基本单元

融合时空注意力的ST-GCN模型的每个基本单元，即图7中的B1、B2、B3等等，均由三个部分共同组成，如图8所示. 即由一个时空注意力模块(图中橘色方块)、一个空间GCN图卷积层(图中蓝色方块)、一个时间TCN卷积层(图中蓝色方块). 数据首先经过时空注意力模块，随后进入空间图卷积模块和时间图卷积模块，分别在空间和时间维度上提取增加了时间注意力和空间注意力后的人体关节点骨架特征信息. 最后，使用残差链接保证训练的稳定性.

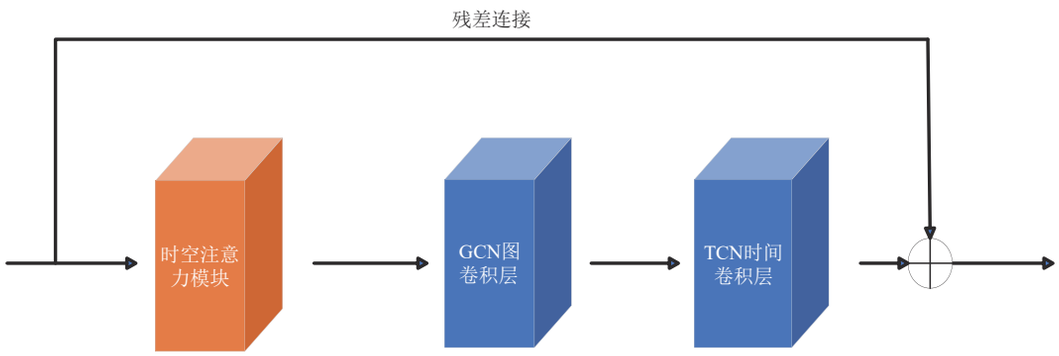

图8　融合时空注意力的改进ST-GCN基本单元

Fig. 8　Basic Unit with spatiotemporal attention of improved ST-GCN

具体计算过程如下：在时间维度上计算注意力权重 W_t，σ 表示sigmoid激活函数，Conv2d代表卷积操作，卷积核大小为(1, 3)，填充padding为(0, 1)：

$$W_t = \sigma(\mathrm{Conv2d}(x)), \tag{3}$$

然后对输入 x 进行时间维度上的加权操作，$\cdot$ 代表相乘：

$$x' = x \cdot W_t, \tag{4}$$

同样地，在空间维度上计算注意力权重 W_s，卷积核大小为(3,1)，填充padding为(1,0)：

$$W_s = \sigma(\mathrm{Conv2d}(x')), \tag{5}$$

然后对输入 x 进行空间维度上的加权操作：

$$x'' = x' \cdot W_s, \tag{6}$$

经过上述处理后的输入特征 x'' 就包含了时间和空间维度上的特征.

3　实验与结果分析

3.1　实验环境和数据集

为验证改进ST-GCN模型在人体动作识别的优越性，对模型分别设置消融实验及对比实验. 实验环境为Ubuntu20.04系统，CPU为Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz，GPU为NVIDIV Tesla P40 24 GB，CUDA版本为11.7，实验环境如表1所示.

选用以下两个公共数据集进行实验：(1)Kinetics数据集[18]包含了约260000个视频剪辑，包含400类动作类别. 该数据集只提供了原始视频剪辑，没有骨架数据，因此需要借用OpenPose[19]工具箱来进行人体关节位置的捕获.(2)NTU RGB+D数据集[20]共包含60个动作类的约56000个动作剪辑. 该数据集有两类划分，CS(Cross-Subject)划分下，一部分志愿者只出现在训练集，一部分只出现在测试集. CV(Cross-View)划分下，用于训练的剪辑来自摄像头2和3，测试集的剪辑来源于摄像头1. 两大公共数据集的训练集和验证集数目如表2所示. 在ST-GCN模型的训练中将两大公共数据集划分为训练集和验证集，并未单独划分测试集，模型结果的评估和测试均是在验证集上进行的.

表1　实验环境

Tab. 1　Experimental environment

实验环境	环境配置
操作系统	Ubuntu20.04服务器
处理器	Intel(R)CPU E5-2630 v4
内存	32.00 GB
显卡	Tesla P40 24 G
框架	Pytorch
硬盘	500 GB

表2　数据集信息

Tab. 2　Data set information

名称	训练集	验证集	总计
Kinetics	240000	20000	260000
NTU RGB+D (CS)	40320	16560	56880
NTU RGB+D (CV)	37920	18960	56880

人体关节点主要有两种表示形式，即18个关节点和25个关节点. 具体划分如图9所示.

3.2 实验评估标准

本文采用的评价标准是Top-n. Top-n指代表输出预测标签最靠前的n类中,有符合人工标注的结果的数目并将结果转化成概率型.举例来说,人工标注标签为A,预测结果为A、B、C、D、E,则Top-1、Top-5都是100%,因为标注标签A出现在预测结果的第1位和前5位;若输出的预测结果为B、A、C、D、E,则Top-1为0%,Top-5为100%,因为标注标签A没有出现在预测结果的第一位,但前5位结果中仍然包含A. Top-2、Top-3、Top-4、Top-n的计算以此类推.ST-GCN使用的是Top-1和Top-5.

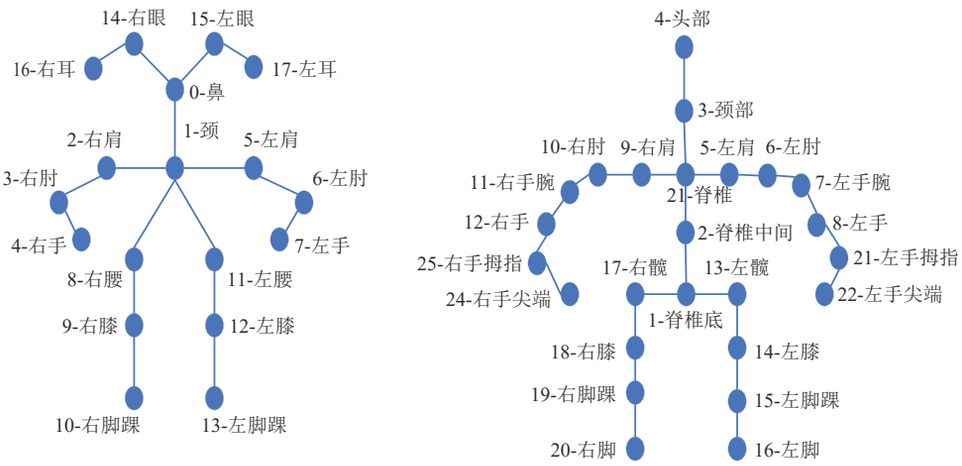

图9 18关节点划分(左)和25关节点划分(右)

Fig. 9 18 joint division (left) and 25 joint division (right)

Top-1和Top-5的计算公式如式(7)、(8)所示:

$$\text{Top-1} = \frac{\sum_i^N \sigma(\text{class}_i^{\text{true}} = \text{rank}_1(\text{class}_i^{\text{pred}}))}{N}, \quad (7)$$

$$\text{Top-5} = \frac{\sum_i^N \sigma(\text{class}_i^{\text{true}} = \text{rank}_5(\text{class}_i^{\text{pred}}))}{N}, \quad (8)$$

其中:N代表样本总数,$\text{class}_i^{\text{true}}$代表第$i$个样本的正确类别,$\text{rank}_1(\text{class}_i^{\text{pred}})$和$\text{rank}_5(\text{class}_i^{\text{pred}})$分别代表第$i$个样本预测概率排名第一的类别和前五的类别.当括号内的条件为True时,$\sigma = 1$,否则$\sigma = 0$. 简而言之,公式上方为正确标签包含在输出的前x个最高分类概率中的个数,下方为总的测试样本数目.

3.3 参数设置

将Kinetics数据集的训练epoch数按原模型作者默认配置设置为50. 考虑到原模型是多卡运行而实验环境为单卡,故将batch_size大小从原来的256设置为64,test_batch_size同样设置为设置为64.base_lr设置为0.1,衰减率设置为0.0001,每10个epoch衰减一次. 而NTU-RGB+D数据集的训练epoch数同样按原模型作者默认配置设置为80,batch_size大小从原来的64设置为32,base_lr设置为0.1,衰减率设置为0.0001,第10个和第50个epoch时衰减. SGD被用来在优化过程中自动调整学习率. 利用Dropout来减轻过度拟合,Dropout设置为0.5.

3.4 对比实验

在两大公共数据集NTU-RGB+D和Kinetics上进行实验,与传统方法如Lie Group、Feature Enc等,CNN方法如Clips+CNN+MTLN等,RNN方法如ST-LSTM、HBRNN、Deep LSTM等,GCN方法如ST-GCN、TCN等进行比较,实验结果见表3和表4. 本文改进模型在两个大型公共数据集上相比于原模型均达到了更高的准确率,这证明了本文方法的有效性.

表3 在NTU RGB+D数据集上的准确率

Tab. 3 Accuracy on NTU-RGB+D dataset /%

方法	CS	CV
Lie Group	50.10	82.80
HBRNN	59.10	64.00
ST-LSTM	69.20	77.70
TCN	74.30	83.10
Clips+CNN+MTLN	79.60	84.80
ST-GCN	81.50	88.30
本文方法	**82.37**	**89.84**

表4 在Kinetics数据集上的准确率

Tab. 4 Accuracy on Kinetics dataset /%

方法	Top-1	Top-5
Feature Enc	14.90	25.80
Deep LSTM	16.40	35.30
TCN	20.30	40.00
ST-GCN	30.70	52.80
本文方法	**31.78**	**54.60**

本文模型在单显卡复现环境下的训练时间和显存占用如表5、表6所示. 其中,表5是在kinetics数

据集上进行的实验结果，表6是在NTU RGB+D数据集上进行的实验结果. 由下表中的实验结果可以看出，在添加新时空注意力模块后，在Kinetics数据集上，改进后模型的每回合所需训练时间和显存占用均大于原模型；但在NTU RGB+D数据集上，二者均小于原模型.

表5　Kinetics数据集上实验结果

Tab. 5　Result on Kinetics dataset

方法	每epoch训练时间(min)	显存占用(MB)
ST-GCN	85	15163
本文模型	140	16604

表6　NTU RGB+D数据集上实验结果

Tab. 6　Result on NTU RGB+D dataset

方法	每epoch训练时间(min)	显存占用(MB)
ST-GCN	36	21819
本文模型	34	18631

由图10可以看出，在Kinetics数据集上，改进后模型的loss值下降比原模型要快，曲线基本在下方.

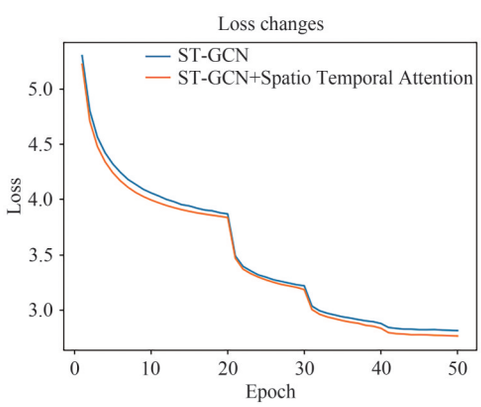

图10　Kinetics数据集上的loss值曲线

Fig. 10　Loss value curve on Kinetics dataset

由图11、图12可以看出：在NTU RGB+D数据集上，无论是在CS划分还是在CV划分下，改进后模型的loss值下降比原模型要快，曲线基本在下方.

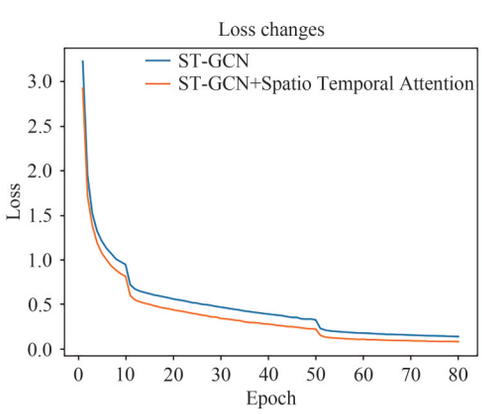

图11　NTU数据集CS划分下的loss值曲线

Fig. 11　Loss value curve of CS divide on NTU RGB+D dataset

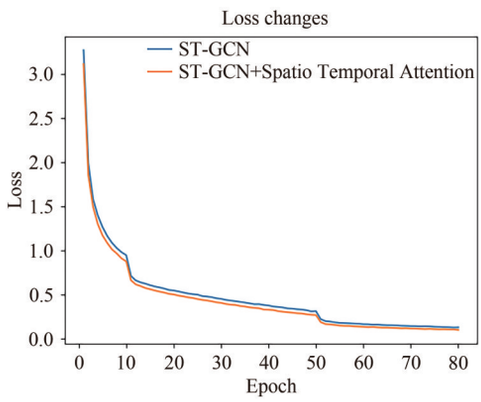

图12　NTU数据集CV划分下的loss值曲线

Fig. 12　Loss value curve of CV divide on NTU RGB+D dataset

3.5　消融实验

本节实验用于验证融合时空注意力机制的实际效果. 将时间注意力、空间注意力、融合时空注意力这三种不同方法分别在两个大型公开数据集上进行实验，使用Top-1和Top-5作为评价标准. 在Kinetics数据集上的实验结果如表7所示. 加入时间注意力后Top-1达到了31.3%，Top-5达到了54.1%，

表7　Kinetics数据集消融实验结果(%)

Tab. 7　Results of Kinetics dataset ablation experiment　/%

方法	Top-1	Top-5
ST-GCN	30.70	52.80
ST-GCN+时间注意力	31.31	54.16
ST-GCN+空间注意力	31.53	54.43
ST-GCN+时空注意力	31.78	54.60

分别比原来高0.6个百分点和1.3个百分点；加入空间注意力后Top-1达到了31.5%，Top-5达到了54.4%，分别比原来高0.8个百分点和1.6个百分点；加入融合时空注意力机制后，在Top-1和Top-5的准确率分别提高1.08%和1.8%.

在NTU-RGB+D数据集上进行实验，同样以Top-1作为评估标准. 和Kinetics数据集上的操作一致，将3种不同的注意力分开分别进行实验和对比，结果如表8所示. 加入时间注意力后，在CS上的Top-1指标达到了81.71%，CV上的Top-1指标达到了89.13%，分别比原模型高了0.21个百分点和0.83个百分点；加入空间注意力后，在CS上的Top-1指标达到了81.82%，CV上的Top-1指标达到了89.55%，分别比原模型高了0.32个百分点和1.25个百分点；加入融合时空注意力机制后，在CS和CV上的Top-1的准确率分别提升0.87%和1.54%的精度，达到了82.37%和89.84%. 由此可知，加入任何形式的注意

力机制,实验数据结果均优于原本的ST-GCN模型.

表8 NTU-RGB+D数据集消融实验结果

Tab. 8 Results of NTU-RGB+D dataset ablation experiment /%

方法	CS	CV
ST-GCN	81.50	88.30
ST-GCN+时间注意力	81.71	89.13
ST-GCN+空间注意力	81.82	89.55
ST-GCN+时空注意力	82.37	89.84

由上述实验结果可知,在加入融合时空注意力模块后,模型的识别准确率相较于未加入注意力模块之前均有提升.融合时空注意力后模型能更为有效地关注人体关节点在运动中的时空联系,更精确地构建动作的全局特征模型.

图13、图14、图15展示的是本文模型的实际识别效果图,四宫格图左上角是输入的原始视频,右上角是对原始视频中的人进行姿态估计的效果,左下角输出的是人体骨骼点的识别和最终动作识别的类别信息,右下角则是将人体骨骼点叠加在原始图像上的效果.

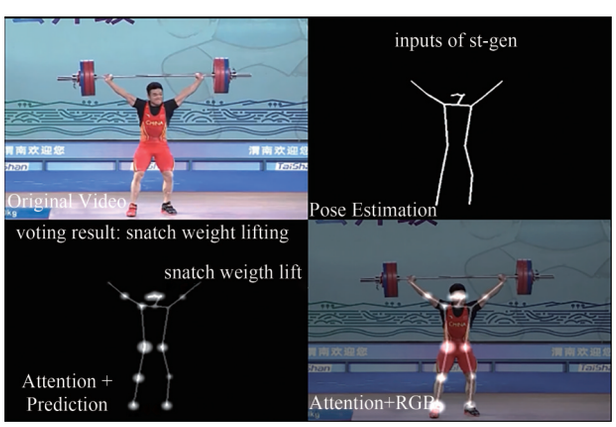

图13 举重动作识别效果

Fig. 13 Action recognition result of weightlifting

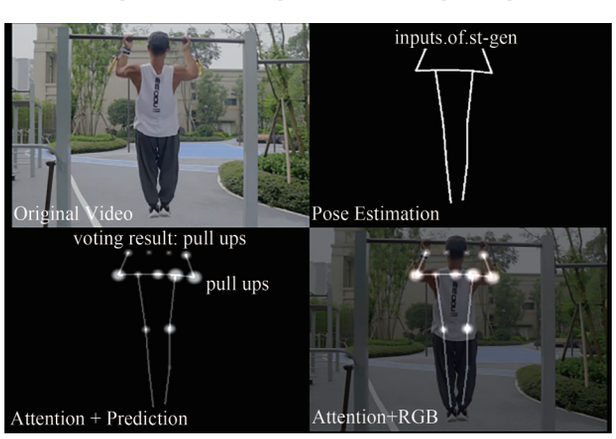

图14 引体向上动作识别效果

Fig. 14 Action recognition result of pull-up

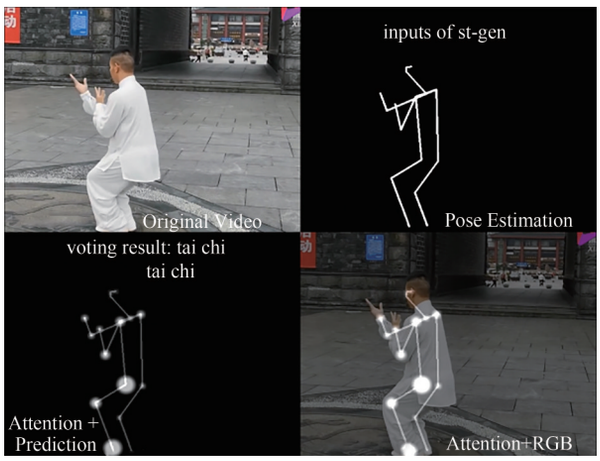

图15 太极动作识别效果

Fig. 15 Action recognition result of taichi

4 结语

针对动作识别模型在训练、学习人体骨架数据时存在的识别精度不高的问题,本文基于时空图卷积网络ST-GCN,提出了一种在原模型基础上加入时空注意力的改进网络模型.引入时空注意力后,可以使网络在提取时域特征和空域特征时有更强的表达能力.实验证明:本文的模型在NTU-RGB+D数据集上的CS标准下取得了82.37%的精度,在CV标准下取得89.84%的精度,相比原来的ST-GCN算法,分别提升0.87%和1.54%的精度.在Kinetics数据集上,本文算法在Top-1和Top-5准确率上取得了31.78%和54.6%的精度,分别提高了1.08%和1.8%.由此验证了本文改进方法的可行性.

参 考 文 献

[1] 马骁,闫育东.基于多尺度时空特征的篮球场景中人体姿态估计[J].中南民族大学学报(自然科学版),2023,42(1):95-102.

[2] FU L, ZHANG J, HUANG K. Beyond tree structure models: A new occlusion aware graphical model for human pose estimation[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 1976-1984.

[3] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 588-595.

[4] FERNANDO B, GAVVES E, JOSE ORAMAS M, et al. Modeling video evolution for action recognition[C]//2015 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR). Boston: IEEE, 2015: 5378-5387.

[5] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]// Computer Vision - ECCV 2016. Amsterdam: Springer, 2016: 816-833.

[6] KE Q, BENNAMOUN M, AN S, et al. A new representation of skeleton sequences for 3D action recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 4570-4579.

[7] LI C, ZHONG Q, XIE D, et al. Skeleton-based action recognition with convolutional neural networks[C]//2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong: IEEE, 2017: 597-600.

[8] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 1110-1118.

[9] SONG S, LAN C, XING J, et al. An end-to-end spa-tio-temporal attention model for human action recognition from skeleton data[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 4263-4270

[10] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks[J]. arXiv Preprint arXiv: 1810.00826, 2018.

[11] ZHANG M, CHEN Y. Link prediction based on graph neural networks [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: ACM, 2018: 5171-5181.

[12] QI S, WANG W, JIA B, et al. Learning human-object interactions by graph parsing neural networks [C]// Proceedings of the 2018 European Conference on Computer Vision. Munich: Springer, 2018: 407-423.

[13] LI R, TAPASWI M, LIAO R, et al. Situation recognition with graph neural networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 4183-4192.

[14] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [J]. arXiv Preprint arXiv: 1609.02907, 2016.

[15] SIMONOVSKY M, KOMODAKIS N. Dynamic edge-conditioned filters in convolutional neural networks on graphs[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 29-38.

[16] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional network for skeleton-based action recognition [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 7444-7452

[17] 徐广, 吴星辰. 基于LSA-HRnet网络的人体姿态估计方法在太极拳运动中的应用[J]. 中南民族大学学报(自然科学版), 2023, 42(6): 839-845.

[18] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv Preprint arXiv: 1705.06950, 2017.

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

[20] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 1010-1019.

(责编&校对　刘钊)