doi:10.19306/j.cnki.2095-8110.2022.03.017

基于相对骨骼点特征和时序自适应感受野的动作识别方法

胡　昊[1],史天运[2],宋永红[3],余　淮[4]

(1. 中国铁道科学研究院研究生部,北京 100081;
2. 中国铁道科学研究院集团有限公司,北京 100081;
3. 西安交通大学软件学院,西安 710049;
4. 中国铁道科学研究院集团有限公司通信信号研究所,北京 100081)

摘　要:针对长时间动作识别难以充分利用时空域信息的问题,提出了基于相对骨骼点特征和时序自适应感受野的动作识别方法。首先,该方法在特征获取部分增加了相对骨骼点特征,以满足节点多样性和互补性要求,将其分别输入到空域图卷积网络,获得空间中相邻关节聚合的局部特征。然后,设计了一个时序自适应感受野网络,以获取在时域中关节变化的局部特征,并且增加了网络对不同持续时长动作的适应性。最后,经过决策级融合模块,计算类别概率,得到分类结果。仿真结果表明,基于 NTU RGB+D 和 Kinetics-skeleton 两大基准数据集,对比多种主流方法,均取得了更高的识别准确率,分别为 96.2%与 60.1%。该方法可以较好地提取不同动作的区别性时间特征,提高了动作时空特征的判别能力。

关键词:动作识别;时序特征提取;图卷积网络;相对骨骼点特征;时序自适应
中图分类号:U495;V249　　**文献标志码**:A　　**文章编号**:2095-8110(2022)03-0132-08

Action Recognition Based on Relative Skeleton Point Features and Temporal Adaptive Receptive Field

HU Hao[1], SHI Tian-yun[2], SONG Yong-hong[3], YU Huai[4]

(1. Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China;
2. China Academy of Railway Sciences Corporation Limited, Beijing 100081, China;
3. School of Software, Xi'an Jiaotong University, Xi'an 710049, China; 4. Signal & Communication Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: Aiming at the problem that it is difficult to make full use of temporal and spatial domain information for long-term action recognition, an action recognition method based on relative skeleton point features and temporal adaptive receptive field is proposed. Firstly, this method adds relative skeleton point features in the feature acquisition part to meet the node diversity and complementarity requirements, and respectively inputs them into the spatial graph convolution network to obtain the local features of the adjacent joint aggregation in space. Then, a time-series adaptive receptive field network is designed to obtain the local features of joint changes in the time domain,

收稿日期:2022-03-01;**修订日期**:2022-03-19
基金项目:中国国家铁路集团有限公司系统性重大课题(P2020T002)
作者简介:胡昊(1991-),男,博士研究生,主要从事模式识别与智能系统方面的研究。
通信作者:史天运(1967-),男,研究员,博士生导师,主要从事智能铁路方面的研究。

and to increase the adaptability of the network to actions of different durations. Finally, through the decision-level fusion module, the category probability is calculated to obtain the classification result. The simulation results show that based on the two benchmark datasets of NTU RGB+D and Kinetics-skeleton, compared with other mainstream methods, this method has achieved higher recognition accuracies, which are 96.2% and 60.1%, respectively. The method can better extract the discriminative temporal features of different actions, and improve the discrimination ability of spatiotemporal action features.

Key words: Action recognition; Temporal feature extraction; Graph convolutional network; Relative skeleton point features; Temporal adaptation

0　引言

动作识别的目标是分析一段视频，判断这段视频中的人所做动作并正确地划分到所属的类别中，在视频理解中有着不可忽视的实际应用价值。动作识别与人们的日常生活息息相关，例如安防监控、人机互动等。传统的动作识别大多基于外观和光流建模，容易受到光线变化、视频中背景等因素干扰，识别精度不是很高。与传统方法相比，基于人体关节点信息的动作识别由于不容易受到背景、光线等因素干扰，通常能表征重要信息。因此，针对基于人体关节点数据的动作识别研究十分必要。近年来，动作识别作为当前识别领域的研究热点，国内外众多研究学者对其展开研究和探索，并且获得了显著的成果。基于人体关节点信息的动作识别方法可以分为以下两类：基于手工特征的方法和基于深度学习的方法。

1）基于手工特征的方法一般不利用深度网络，以人工提取特征对人体骨骼的空间和时间进行动态建模，然后用分类器进行人体动作的识别。这些人工提取到的特征包含了对身体部位的旋转和平移方法[1]，以及关节轨迹的协方差矩阵方法[2]等。另外，Xia L. 等[3]提出了用三维关节位置的直方图代表骨架序列里面的每一帧，从而进行时间动态建模。但是，这种方法和基于手工特征的视频动作识别方法的缺点一样，手工特征的设计和计算极为复杂，且没有办法全面地表征动作在时序上的演化，使得最终动作识别的性能不理想。

2）基于深度学习的方法由于在计算量、类脑计算方式等方面优于传统的手工特征方法，结果相对更好。基于深度学习的方法主要有三种框架：基于循环神经网络（Recurrent Neural Network，RNN）的方法、基于图像的方法和基于图卷积的方法。

基于循环神经网络的方法主要是为了获取长时间的时序信息，将骨架数据表示为关节序列，然后用基于循环神经网络改进的长短期记忆（Long Short-Term Memory，LSTM）神经网络对其建模[4-7]，为动作识别找到样本序列中信息最为丰富的帧，并通过关键帧帮助网络进行分类，从而提高识别精度。Song S. 等[8]引入了一个时空注意模型，用注意力机制为视频中不同的帧和节点分配不同的权重，但训练过程很复杂。A. Jain 等[9]主要对骨骼三个部位的关系进行建模，包括脊柱、手臂和腿，网络架构是将循环神经网络与图结合在一起。Du Y. 等[10]设计了一种通过级联方式组合人体骨骼各个部位的方法，用长短期记忆网络进行建模时序运动。但是，基于循环神经网络的动作识别方法也有缺点，它主要考虑时序特征，在空间位置信息的获取方面稍有不足，且网络相对复杂，没有办法加深网络。

基于图像的方法主要是把骨架三维坐标表示为特殊的图片，也可以称为伪图像，然后用卷积网络对图片进行特征提取和训练。Ke Q. 等[11]提出了一种新的三维骨架序列表示方法，即将样本序列中的柱坐标（3 个坐标表示骨骼节点位置）转换成伪图像（3 个灰度图像），然后再利用深度卷积网络进行训练和时空特征的学习。前者是转换成 3 段灰度图像，而 Liu M. 等[12]是将序列转换成一系列的彩色图像并输入到卷积神经网络（Convolutional Neural Network，CNN）中进行特征获取，最终进行动作识别。

基于图卷积的方法将数据建模成以骨骼关节为顶点、以骨骼边为边的图，并通过卷积学习图中不同节点之间的信息交流，从而得到图中每个顶点的嵌入特征表示。Yan S. 等[13]将图卷积神经网络引入骨骼动作识别中，设计了时空图卷积网络，用于训练样本序列，最终得到动作识别结果。这种方法借鉴了将 3D 分解为 2+1D 的思想，通过空域上

的图卷积提取空间信息，并通过在时域上提取相邻帧卷积之后的特征来提取时序信息，从而通过聚合空域时域信息来捕捉时空的变化关系。Tang Y. 等[14]设计了一个深度渐进强化学习模型，通过时间上的类注意力方法选择最有代表性的帧，也就是提取含有大信息量的帧，并去除含有一些无用信息的帧，然后输入到图卷积网络中进行训练。Zhang X. 等[15]提出了给骨头边卷积的思想，用双流卷积网络分别卷积骨头节点和骨头边，结合两个网络得到最终结果，提升了准确率。Shi L. 等[16]在时空图卷积的基础上对邻接矩阵策略进行改进，用自注意力机制设计每个样本的邻接矩阵，大大增强了对空间特征的提取。Shi L. 等[17]将时空图卷积从无向图变成有向图，提高了抽取空间特征的有效性。Li M. 等[18]用多个图进行卷积，不仅关注有物理连接的关节点之间的潜在联系，还注重没有骨骼边相连两个节点之间的关系。

动作识别在近些年来受到了广大研究者们的关注，基于人体关节点信息的动作识别方法凭借其对运动速度、背景干扰和摄像机视点的鲁棒性，取得了不错的成绩。但是，现有的基于关节点信息的动作识别方法时空特征判别能力不强，具体表现在固定单一核的时间卷积无法为不同动作获得更有区别性的时间特征，对视频中持续时间长的动作类别识别效果难以保证，导致识别精度有所影响。针对该问题，提出了基于相对骨骼点特征和时序自适应感受野的动作识别方法，可以较好地提取时空特征。实验结果表明，对比其他方法，该方法在基准数据集上获得了识别性能的提升。

1 算法框架概述

现有方法大多只关注人体骨骼关节点在时间上的位移，而人体关节点的空间相对位置信息等特征在基于骨骼信息的动作识别中也起着很重要的作用，但这些特征往往都会被忽略，而且多种特征之间具有互补性和多样性。另一方面，神经网络想要提高泛化能力，需要大量的数据来支撑。而基于骨骼信息的动作识别，输入的数据是关节序列的三维坐标，一帧中的骨骼节点太少会出现过拟合的情况，导致训练出的结果精度不高。

因此，本文提出了基于相对骨骼点特征和时序自适应感受野的动作识别方法，可以较好地解决现有方法无法为不同的动作获得更有区别性的时间特征的问题，提升对视频中持续时间长的动作类别的识别准确率，整体架构如图 1 所示。

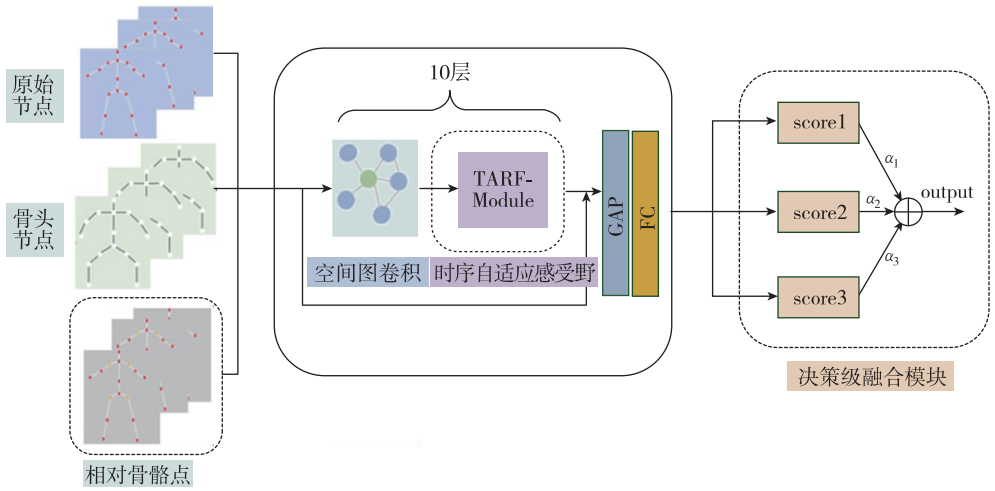

图 1 算法整体框架

Fig. 1 Overall framework of the algorithm

首先，在特征获取部分根据原始三维节点特征，计算其输入特征，分别得到骨头特征和相对骨骼点特征，以满足节点多样性和互补性要求。让输入特征分别经过 10 个时空网络块进行训练，每个块包含了一个空域图卷积网络和时序自适应感受野模块。通过空域图卷积网络训练，得到空间中相邻关节聚合的局部特征，再经过时序自适应选择不同感受野的信息，获取在时域上关节变化的局部特征，增强了网络对不同持续时长动作的适应性。之后，用残差相加的方法，将训练后的特征与原始特

征结合，在 10 层网络训练之后，经过决策级融合模块，计算 softmax 层的类别概率，通过熵权法求得每个特征流的融合权重，从而得到分类结果。

2　算法模块设计

2.1　相对骨骼点特征

骨骼关节点的空间相对位置信息等特征，在基于骨骼信息的动作识别中具有重要作用，但这些特征往往都会被忽略。神经网络要提高泛化能力，需要大量的数据来支撑，而基于骨骼信息的动作识别输入的数据是关节序列的三维坐标，一帧中的骨骼节点太少，会出现过拟合的情况，导致训练出的结果精度不高，而且多种特征之间具有多样性和互补性。三维骨架序列是一个五维的特征向量 $[N,C,T,V,M]$。其中，N 是指批量大小；T 是指每个样本的帧的数量；V 是指人体骨架的关节数；M 是指人的数量；C 是指输入特征的通道数量，包含了数据集从深度摄像机中采集的原始人体骨骼三维节点 $v_{i,t}=(x_{i,t},y_{i,t},z_{i,t})$，其中 i 表示关节点序号，t 表示帧序号。由于人在做动作时，人体关节点有可能重合在一起，导致动作的误判。骨头特征指的是人体骨骼中的骨骼边，加入骨头特征可以一定程度上解决这个问题。因此，骨头特征也被广泛用在基于关节点信息的动作识别中。

但是，当人体动作发生时，骨骼特征仅计算了存在物理骨骼边关节点间的空间关系信息，这是不够的。当两个关节点之间不存在骨骼边时，这些关节点之间的空间关系也很重要。例如，抹脸和梳头这两个动作，都是手部节点与头部节点距离很近，这些节点没有骨骼边连接，如果计算这些关节之间的相对位置信息，会发现特征向量的大小和方向都不一样。因此，借鉴 Ke Q. 等[11]用三维骨架序列生成图像帧中处理序列的方法，再结合关节点之间的相对位置信息，提出了相对骨骼点特征。

为防止冗余信息的产生，在同一帧内仅选几个节点，用这几个节点的位置信息当作源关节坐标点，计算其他节点针对这 4 个节点的空间相对位置信息。源关节坐标点选取的条件一是要反映出其他节点的位置信息，以及与大多关节点的潜在联系；二是在做动作时可以保持稳定的状态，如图 2 所示，有颜色的 7 个关节点可以作为源关节坐标点，分别是右肩节点、左肩节点、最中间的 3 个脊柱节点、右臀节点和左臀节点。但是脊柱节点离其他 4 个节点的距离很近，如果都选为源关节坐标点，会导致信息的重复，所以仅以图中黄颜色的右肩节点、左肩节点、右臀节点和左臀节点为源关节坐标点。

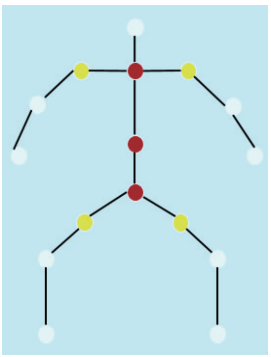

图 2　源关节坐标点的筛选

Fig. 2　Screening of source joint coordinate points

相对骨骼点特征的计算方法是以其他关节点为目标关节点，分别计算目标关节与 4 个源关节的差值，得到 4 个特征向量，再将这 4 个特征向量在通道维进行拼接，输入到网络中进行训练。

具体的计算公式如式(1)、式(2)

$$J_{\mathrm{rej}}=\{(v_{i,t}-v_{j,t})\mid i\in V,t\in T,j\in V'\} \tag{1}$$

$$J_{\mathrm{re}}=J_{\mathrm{re1}}\oplus J_{\mathrm{re2}}\oplus J_{\mathrm{re3}}\oplus J_{\mathrm{re4}} \tag{2}$$

式(1)中，$v_{i,t}$ 表示目标关节点坐标；$v_{j,t}$ 表示源关节点坐标；V 表示人的骨架的关节点集合；V' 表示 4 个源关节点坐标集合；T 表示帧数。式(2)表示将计算得到的 4 个特征向量拼接在一起，生成相对骨骼点特征。

2.2　时空网络模块

时空网络块如图 3 所示，包含了一个空间图卷积网络和时序自适应感受野模块。Conv-s 就是图 1 所示的空间图卷积网络，将特征输入后可以得到空间中相邻关节聚合的局部特征。Conv-t 即图 1 所示的时序自适应感受野模块，通过自适应选择不同感受野的信息，获取在时域上关节变化的局部特征。为了获得更好的时空特征，在获得空间特征后，经过批标准层，加快收敛速度，之后加入残差模块，稳定特征训练过程，通过 relu 操作增强各层之间的非线性关系，减少过拟合，再将经过这些层处理后的空间特征输入到时域中进行卷积，得到时空卷积。

特征要经过 10 个时空网络块进行训练，这 10 层网络的配置为第 1 层的输入通道数为原始节点的通道数，1～4 层的输出通道为 64，5～7 层的输出通

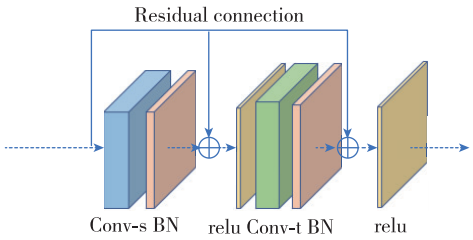

图 3 时空网络块框架

Fig. 3 The spatiotemporal network block framework

道为 128，8～10 层的输出通道为 256，第 5 层和第 8 层步长设置为 2，等同于池化层。

2.3 时序自适应感受野模块

在动作识别领域，大多数方法选择在空域图卷积网络进行改进，但在时域大多用单一固定的时间卷积层（Temporal Convolutional Nets，TCN）对特征进行提取。这样提取出的特征不足以为不同的动作获得更有区别性的时间特征，对视频中持续时间长的动作类别识别效果难以保证，导致识别精度有所影响。因此，考虑用非线性方法，允许每个神经元根据上一层的多个感受野尺度自主选择不同分支的卷积层信息。时序自适应感受野模块的主要原理是计算不同感受野通道的注意力权重，使网络自适应获取不同感受野的信息。

时序自适应感受野模块的网络结构如图 4 所示。对输入特征分为四路卷积，4 个分支的卷积核分别为 3×1、5×1、7×1 和 9×1，分别得到 4 个特征，对这 4 个特征进行简单的像素级相加融合，得到特征 U_s。为了建模通道之间的依赖关系，对特征 U_s 沿着 T 和 V 维度求平均值，从而得到每个通道的信息。之后，为了完成针对跨通道信息的提取，用具有自适应卷积核的快速一维卷积进行 4 次快速一维卷积，得到 4 个 c 维通道的特征向量，然后用 softmax 进行归一化。用通道间的注意力方法以自适应选择 4 个分支的信息，得到权重矩阵注意力向量，再用权重矩阵对 U_1、U_2、U_3 和 U_4 进行加权操作并求和，得到最后特征 Fea_V。这样，最后网络融合了不同感受野的信息，且不会造成信息的冗余。

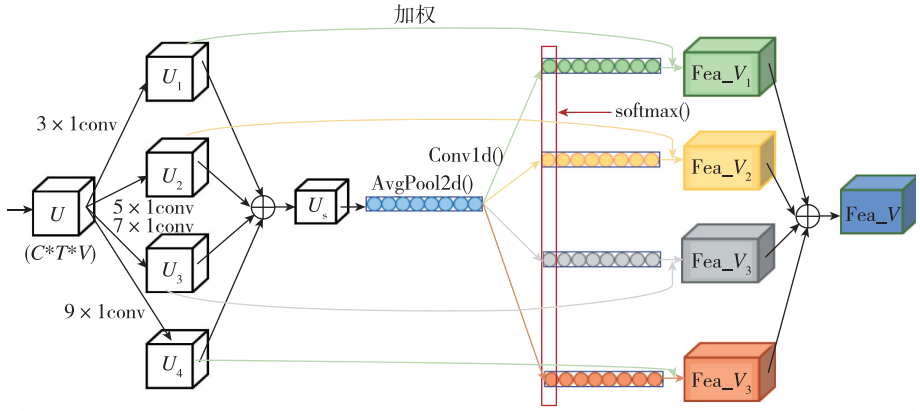

图 4 时序自适应感受野模块

Fig. 4 Temporal adaptive receptive field module

2.4 决策融合

由于利用特征融合的方法可以实现多种特征之间的互补性和多样性，综合隐含在不同特征中的信息，能够防止过拟合。因此，选择用决策级融合的方法如采用多流网络结构，将各特征输入到共享网络层中，再将 softmax 层的最后分数加起来，最后分数最高的那一类就是训练后的分类结果。

由于各个特征流占的比重不同，不能简单地用全是 1 的参数来融合，因此，用熵权法确定各个流融合的分数之前的权重。根据熵值判断各个特征流通过网络得出分数的离散程度，也就是根据各个特征和它对应值的变异性大小来确定客观权重，在这里离散程度越大，该特征流对综合评价的影响越大。

3 实验结果分析

3.1 实验数据及评价方法

本文提出的动作识别方法主要在 NTU RGB+D 数据集和 Kinetics-skeleton 数据集进行实验，在 NTU 数据集上进行消融实验，之后分别基于两个数据集上的实验结果，与其他方法进行对比分析。

准确率的计算如式(3)所示

$$Acc=\frac{right_{\text{sample}}}{all_{\text{sample}}}\times 100\% \tag{3}$$

式中，$right_{\text{sample}}$ 是正确分类的动作序列样本数；all_{sample} 是全部动作序列样本数。

3.2　参数设置

本文提出的算法在两个数据集上批量大小都设置为 32，权重衰减都设置为 0.0001，初始学习率设置为 0.1，用随机递度下降(Stochastic Gradient Descent，SGD)算法进行优化。在 NTU RGB+D60 数据集上进行 50 次迭代，用 MultiStepLR()函数在第 30、40 次迭代进行学习率的调整，超参数 gamma 设置为 0.1，调整学习率时将学习率除以 10。该方法基于图形处理器(Graphics Processing Unit，GPU)进行训练和测试。在 Kinetics-skeleton 数据集上进行 65 次迭代，在第 45、55 次迭代进行学习率的调整，超参数 gamma 设置为 0.1，调整学习率时将学习率除以 10。

3.3　实验结果与分析

本文提出的动作识别方法主要在 NTU 数据集上做消融实验，设计了消融实验来验证每个模块的有效性。在 NTU RGB+D60 数据集上和 Kinetics-skeleton 数据集上与当前基于骨骼信息的图卷积动作识别方法进行比较，验证了该方法可以有效提高动作识别的精度。

1)基于三维骨架特征的实验结果

为了验证相对骨骼点特征的有效性，分别基于各特征在 NTU RGB+D60 数据集上进行实验。单独输入 3 个特征进行结果比对。分别将原始骨骼节点、骨头特征和相对骨骼点特征输入到设计好的网络中，得到对 60 类动作的识别结果，如表 1 所示。

表 1　基于各个三维骨架特征在 NTU RGB+D60 数据集上的识别结果

Tab. 1　Recognition results on NTU RGB+D60 dataset based on each 3D skeleton feature

三维骨架特征	CS(top-1)	CS(top-5)	CV(top-1)	CV(top-5)
原始骨骼节点	86.94%	97.48%	94.34%	99.08%
骨头特征	87.36%	97.73%	94.04%	99.17%
相对骨骼点特征	**87.57%**	**97.75%**	**94.14%**	**99.18%**

其中，CS 指跨主题评价指标；CV 指跨视角评价指标；top-1 准确率计算的是预测结果中概率最大的正确类样本数/总样本数；top-5 计算的是预测结果中概率最大前五名中正确类的样本数/总样本数。由于跨视角评价指标的准确率已经很高，此时，主要从跨主题的角度对结果进行分析。从表 1 中可以看到，骨头特征在跨主题的评价指标上准确率要高 0.42%，这是因为骨头特征是将人的骨骼边输入到网络中，解决了不同动作骨头节点可能重合造成误判的问题。用相对骨骼点特征进行动作识别，在跨主题的评价指标上，准确率要比基于原始节点的准确率高 0.63%。考虑到两个关节点之间不存在骨骼边，用相对骨骼点可以计算出这些关节点之间的空间关系。例如，抹脸和梳头这两个动作，都是手部节点与头部节点距离很近，这些节点是没有骨骼边连接的，如果计算这些关之间的相对位置信息，会发现特征向量的大小和方向都是不一样的。

2)时序自适应感受野模块实验结果

为了验证该模块对网络带来的提升效果，对比了以原始节点作为输入特征的基线识别动作的准确率与设计本模块之后识别动作的准确率，如表 2 所示。

表 2　基于原始节点的基线和提出模块后的识别结果的对比

Tab. 2　Comparison of recognition results based on original node in the baseline and the proposed module

方法	CS(top-1)	CS(top-5)	CV(top-1)	CV(top-5)
基线方法	85.82%	97.13%	93.54%	98.84%
时序自适应感受野模块	**86.94%**	**97.48%**	**94.34%**	**99.08%**

基线网络经过训练后，在跨主题评价指标上的准确率是 85.83%，在跨视角评价指标上的准确率为 93.54%；而在网络中设计了时序自适应感受野模块之后，在跨主题评价指标上的准确率是 86.94%，精度提高了 1.11%，在跨视角评价指标上的准确率为 94.34%，精度提高了 0.8%。精度的提高证明了设计模块的有效性。时序自适应感受野模块从根本上主要解决的是在所有动作序列中，有的动作关键性阶段持续时间很短，有的动作关键性阶段持续时间长，即提取到的特征不足以覆盖每个动作所有关键阶段的问题。分别查看关键性阶段持续时间很短和很长的动作分类的准确率，关键性阶段持续很短的动作可以用读书这个类(3 帧有个明显的翻页动作)进行结果验证，关键性阶段持续很长的动作可以用玩手机、打字这两类(关键性阶段持续大概在 9 帧左右)进行结果验证。

如表 3 所示，基于本文设计的网络，读书动作识

别准确率比基线提高了 4%，玩手机动作比基线提高了 6%，打字动作比基线提高了 5%。这个实验结果可以进一步证明本模块能够有效获取不同时域长短的动作特征。

表 3 基于基线和本文网络比对动作类的分类准确率

Tab. 3 Comparison of classification accuracy of action classes based on baseline and the proposed module

方法	读书(A11)	玩手机(A29)	打字(A30)
基线方法	52%	60%	64%
时序自适应感受野模块	56%	66%	69%

3)与其他方法比较的实验结果

在 NTU RGB+D60 与 Kinetics-skeleton 两个数据集上，将本文方法与其他基于骨骼信息的图卷积动作识别方法进行比较。时空图卷积网络(Spatio-Temporal Graph Convolutional Networks，STGCN)[13]方法用适合的图卷积网络提取空间特征；动作结构图卷积网络(Actional-Structural Graph Convolutional Networks，ASGCN)[18]方法用多个图进行卷积，注重没有骨骼边相连的两个节点之间的关系；双流自适应图卷积网络(Two-Stream Adaptive Graph Convolutional Networks，2S-AGCN)[16]方法在时空图卷积的基础上对邻接矩阵策略进行了改进，增强了对空间特征的提取；有向图神经网络(Directed Graph Neural Networks，DGNN)[17]方法将时空图卷积从无向图变成有向图，提高了抽取空间特征的有效性。

从表 4 的精度对比实验可以看出，本文方法对动作识别的效果是有提升的。跨主题评价指标的准确率(top-1)比前四种方法中最高的识别效果高出 0.3%左右。跨视角指标的准确率更高一点，说明基于骨骼信息对视角差异是有鲁棒性的。另外，在跨主题评价指标上的较高准确率，说明本文提出的方法可以较好地提取到时空特征。

表 4 本文方法在 NTU RGB+D60 数据集上与当前多种方法的识别结果对比

Tab. 4 Comparison of recognition results of our method and the current methods on the NTU RGB+D60 dataset

评价指标	STGCN	ASGCN	2SGCN	DGCN	本文
跨主题评价	81.5%	86.8%	88.1%	89.9%	90.2%
跨视角评价	88.3%	94.2%	94.7%	96.1%	96.2%

从表 5 的精度对比实验可以看出，Kinetics-skeleton 数据集上比前四种方法中最高的识别效果高出 0.5%(top-1)左右。在两个数据集上的精度提升说明了本文方法的有效性，因为在特征获取部分增加了相对骨骼点特征，满足了节点多样性和互补性要求；并且通过时序自适应感受野网络，获取了在不同时域上关节变化的局部特征，较好地解决了现有方法无法为不同动作获得更有区别性的时间特征的问题。

表 5 本文方法在 Kinetics-skeleton 数据集上与当前多种方法的识别结果对比

Tab. 5 Comparison of recognition results of our method and the current methods on the Kinetics-skeleton dataset

评价指标	STGCN	ASGCN	2SGCN	DGCN	本文
Top-1	30.7%	34.8%	36.1%	36.9%	37.4%
Top-5	52.8%	56.5%	58.7%	59.6%	60.1%

为了验证本文方法在实际应用场景中的有效性，在基于火车站与铁路沿线综合监控视频录像中，构建了包含奔跑、跌倒、攀爬、抽烟和行走等动作类别的数据集，同时也从网络视频中选取了部分同类别数据补充到其中，数据集共 86 段视频，5 大类动作类别。本文在该数据集上与其他动作识别方法进行比较。STGCN、ASGCN、2SGCN、DGCN 与本文方法的识别精度分别为 71.3%、72.1%、71.9%、73.3%和 78.6%，可见本文方法在实际场景中的动作识别能力相比其他方法取得了提升。

4 结论

本文针对长时域动作识别率较低的难题，提出了基于相对骨骼点特征和时序自适应感受野的动作识别方法，算法分析与实验结果表明：

1)在 NTU RGB+D 数据集进行了消融实验，相对骨骼点特征在跨主题评价指标与跨视角评价指标上均取得了最高的准确率，在 CS 上取得了最高准确率提升，证明相对骨骼点特征能有效提取关节点之间的空间关系。

2)在时序自适应感受野模块上，对比基准方法均取得了最高的准确率，特别是通过关键性阶段持续时间较长的动作如玩手机，验证了该模块在时域特征提取上的有效性。

3)在基准数据集 NTU RGB+D 和 Kinetics-skeleton 上，对比了 STGCN、ASGCN 及 2S-AGCN 等多种主流方法，均取得了最高的识别率，说明整体方法能够较好地提取不同动作的区别性时间特征，提高了

动作识别能力。

在实际场景数据中,对比多种主流方法,该方法也取得了最高的识别率。可见在不同的数据集上,该方法能够更好地提取不同动作的时空特征,具有较强的实用价值。

参考文献

[1] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014: 588-595.

[2] Hussein M E, Torki M, Gowayyed M A, et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations[C]// Proceedings of 23th International Joint Conference on Artificial Intelligence, 2013.

[3] Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3D joints [C]// Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012: 20-27.

[4] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 1010-1019.

[5] Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]// Proceedings of European Conference on Computer Vision. Springer, Cham, 2016: 816-833.

[6] Zhang P, Lan C, Xing J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]// Proceedings of IEEE International Conference on Computer Vision. IEEE, 2017: 2117-2126.

[7] Li S, Li W, Cook C, et al. Independently recurrent neural network (IndRNN): building a longer and deeper RNN[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 5457-5466.

[8] Song S, Lan C, Xing J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]// Proceedings of AAAI Conference on Artificial Intelligence, 2017: 4263-4270.

[9] Jain A, Zamir A R, Savarese S, et al. Structural-RNN: deep learning on spatio-temporal graphs[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 5308-5317.

[10] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015: 1110-1118.

[11] Ke Q, Bennamoun M, An S, et al. A new representation of skeleton sequences for 3D action recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 3288-3297.

[12] Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition, 2017, 68: 346-362.

[13] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Proceedings of AAAI Conference on Artificial Intelligence, 2018.

[14] Tang Y, Tian Y, Lu J, et al. Deep progressive reinforcement learning for skeleton-based action recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 5323-5332.

[15] Zhang X, Xu C, Tian X, et al. Graph edge convolutional neural networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(8): 3047-3060.

[16] Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 12026-12035.

[17] Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with directed graph neural networks[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 7912-7921.

[18] Li M, Chen S, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 3595-3603.

(编辑:孟彬)