

语义引导下自适应拓扑推理图卷积网络的人体动作识别

林哲煌, 李东

(广东工业大学 自动化学院, 广东 广州 510006)

摘要: 图卷积网络(Graph Convolutional Networks, GCN)对于基于骨架关节点信息的人体动作识别任务具有天然的优势, 越来越受到重视。图卷积网络的关键在于如何获取更丰富的特征信息以及采用更合理的拓扑结构。本文改进了人体骨架关节点及其语义信息(关节点类型和帧索引)的特征融合方式, 集成为一个语义信息编码模块, 从而更适用于复杂的多层网络。在语义信息编码模块的语义引导下, 网络可以获取更丰富的关节点特征信息。其次, 本文提出了一种拓扑结构推理网络, 结合卷积神经网络(Convolutional Neural Networks, CNN)高效的特征学习能力, 自适应地根据不同动作样本的上下文特征信息学习不同的邻接矩阵, 有助于网络摆脱固定拓扑结构的局限性。将上述方法应用于双流自适应图卷积网络, 本文提出了一种语义引导下多流自适应拓扑推理的图卷积网络。实验结果证明, 本文的方法使图卷积网络识别精度有了明显的提高, 在基于骨架信息的人体动作识别大型数据集NTU RGB+D、NTU RGB+D 120上均达到了目前先进水平。

关键词: 动作识别; 图卷积网络; 人体骨架; 邻接矩阵

中图分类号: TP391.4

文献标志码: A

文章编号: 1007-7162(2023)04-0045-08

Semantics-guided Adaptive Topology Inference Graph Convolutional Networks for Skeleton-based Action Recognition

Lin Zhe-huang, Li Dong

(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Graph convolutional networks (GCN), with natural advantages for skeleton-based action recognition, has attracted more and more attention. The key lies in how to obtain richer feature information and the design of the skeleton topology. In this research, the feature fusion method of joint and semantics (joint type and frame index) is improved, and integrated into a Semantics Coding Module (SCM), which is more applicable for complex multi-layer networks. Guided by the SCM, the network can obtain more feature information of skeleton. Secondly, a skeleton Topology Inference Network (TIN) is proposed, which adaptively learns different adjacency matrices according to the context information of different samples with the efficient feature learning ability of CNN, so that the network can get rid of the limitation of fixed topology. By applying the SCM and TIN to 2s-AGCN, we propose a semantics-guided multi-stream adaptive topology inference graph convolutional network for skeleton-based action recognition. Extensive experiments on datasets, NTU RGB+D and NTU RGB+D 120, demonstrate that our methods obviously improve the accuracy of network and our model has achieved the state-of-the-art performance.

Key words: action recognition; graph convolutional network; skeleton; adjacency matrix

人体动作识别, 又称人体行为识别, 历经多年的发展, 为人与机器或人与人之间的远距离自动交互提供了可能, 成为了一项具有重要发展背景的深度学习

任务, 在医学、交通、教育、娱乐、安防监控、元宇宙等领域有着广阔的应用前景。早在深度学习未发展成熟之前, 很多相关研究已经采取了手动标注的三

收稿日期: 2022-06-13

基金项目: 广东省自然科学基金资助项目(2021A1515011867)

作者简介: 林哲煌(1996-), 男, 硕士研究生, 主要研究方向为机器学习、深度学习、动作识别

通信作者: 李东(1983-), 男, 副教授, 博士, 主要研究方向为模式识别、机器学习、人脸识别, 机器视觉, E-mail: dong.li@gdut.edu.cn

维动作识别^[1-2]。传统方法都是将空间信息转换为手工设置的特征值^[3],再利用传统分类方法如支持向量机(Support Vector Machine, SVM)^[4]等对特征进行分类,但这些方法存在局限性,主要表现在特征向量一旦确定就无法改变,不会随着训练过程的进行而适应输入的变化,抗干扰能力较弱^[5]。随着深度学习的不断发展,基于深度神经网络的人体动作识别算法逐步替代传统手工特征算法成为研究主流^[6-7]。

目前基于深度学习的相关算法分为两大技术流派,一个是基于RGB视频数据的动作识别技术^[8-9],另一个是基于人体骨架关节信息的动作识别技术。相比于RGB视频数据信息,人体骨架信息不受背景、光照等噪声的干扰,且骨架数据维度较小、计算效率高,能够提供良好的人体行为表征,因此受到越来越多的关注。

早期的深度学习算法是将人体的骨架关节信息视为一组独立的特征,并将其处理为一种特征序列或者伪图,然后输入到卷积神经网络(Convolutional Neural Networks, CNN)^[10-12]或者循环神经网络(Recurrent Neural Networks, RNN)^[13-15]中处理特征信息进行动作分类。然而人体是天然的拓扑结构,基于CNN或RNN的算法忽略了在运动过程中人体关节之间内在的时空相关性,而关节之间的相关性由人体的拓扑结构表征,是基于人体骨架信息的动作识别不可忽视的重要信息^[16]。图卷积网络(Graph Convolutional Networks, GCN)对于处理人体关节的拓扑结构的特征信息有着巨大的优势,Yan等^[17]提出了ST-GCN,最先将GCN结合时域卷积网络(Temporal Convolutional Networks, TCN)提取关节的时空特征,并应用到人体动作识别中。在此基础上,Shi等^[18]提出了2s-AGCN,该双流网络添加了关节点之间的骨架信息作为特征信息,增加了网络获取的信息量,可以自适应学习针对不同样本不同的拓扑结构,大大提高了动作的识别率。Zhang等^[19]提出了轻量级的网络SGN,首次提出了关节的语义表示,提升了动作识别精度,但其特征融合方式并不适用于复杂的多层网络。

后来的研究大都是以提高GCN计算效率以及特征提取的能力为目标^[20-21]。Li等^[22]提出了轻量级网络PR-GCN,在不影响精度的前提下,大大减少了网络的参数量,同时显著提高了计算效率。Yang等^[23]提出了PGCN-TCA,使用了混合时间和通道注意的方法来提取不同帧和通道的不同重要性。Ding等^[24]提出的Sem-GCN包含了3种类型的图:结构图、动作图和

注意图,增大了图卷积核的感受野,同时也捕捉到关节间的潜在相关性。这些改进后的GCN计算方法^[22-24]增加了网络的灵活性和鲁棒性,以适应不同的类型的动作样本。然而,这些算法忽略了各节点时间空间上语义信息的重要性,在融合更加丰富的时空特征信息,自适应地学习更加有效的拓扑结构来应对复杂的动作方面还有很大的提升空间。

因此,本文改进了SGN原有的语义信息融合方式,将关节类型、帧间索引的语义信息与数据集原有的关节信息特征融合,集成一个语义信息编码模块,再通过残差连接传递到深层的网络。这样的方式更适用于大型的多层网络,语义的引导也为网络提供了更加丰富的关节特征信息。其次,针对现有GCN算法固定的邻接矩阵的局限性,本文提出了一种拓扑结构推理网络,结合CNN特征学习能力,能够自适应地根据不同样本的上下文特征信息而学习到不同的拓扑结构。本文将语义信息编码模块与拓扑结构推理网络应用到双流自适应图卷积网络2s-AGCN中,增加了相关运动流分支,提出了一种语义引导下的自适应拓扑推理图卷积网络,精度得到了显著的提升。本文的语义信息编码模块和拓扑结构推理网络适用于大多数GCN算法。实验结果证明,本文提出的网络在基于骨架信息的人体动作识别大型数据集NTU RGB+D、NTU RGB+D 120上均达到了目前先进水平。

1 本文方法

1.1 整体结构

本文提出的语义引导下自适应拓扑推理图卷积网络总体结构如图1所示。首先,由人体关节间相对空间位置信息计算得到人体骨架的空间位置信息,再分别计算关节和骨架邻帧位置向量差得到关节和骨架的运动信息。这4类信息经过语义信息编码模块,融合关节类型和帧间索引的高级语义信息,然后将编码及融合后得到的关节信息、关节运动信息、骨架信息、骨架运动信息作为多流自适应图卷积网络的各个分支对应的输入,并可通过残差连接传递到更深层的网络。多流自适应图卷积网络的每个分支都由9个自适应图卷积块组成,每个自适应图卷积块又由含拓扑结构推理网络的自适应图卷积层、Batch Normalization层、Relu激活函数、Dropout操作、TCN组成。在进行GCN计算时,通过自适应图卷积块提取空间和时间上的有效特征信息。4个分支独立计算,对各自计算后的

Softmax分数加权求和,作为最终的Softmax分数,选用交叉熵损失函数作为网络最终的损失函数进行

梯度反向传播,从而实现基于骨架关节信息的人体动作识别。

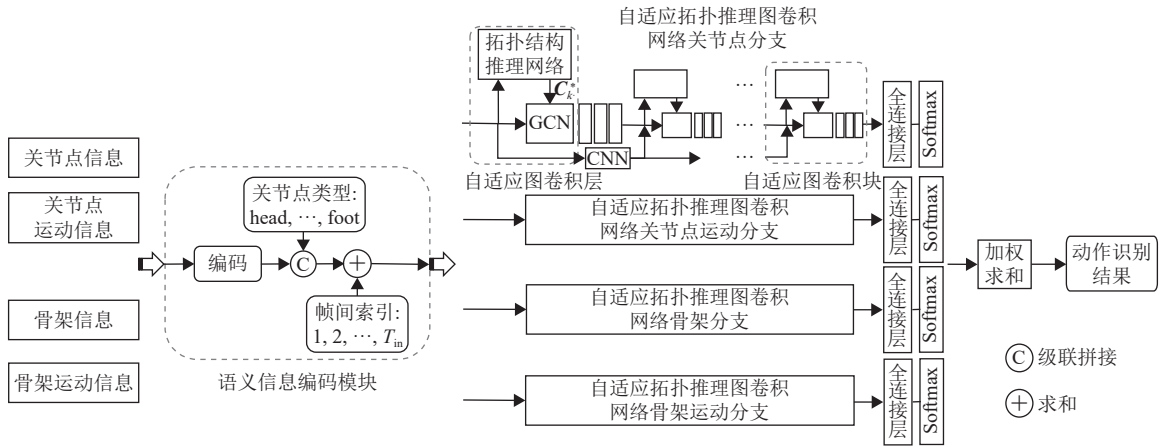


图1 本文网络整体结构

Fig.1 The overview of proposed network

1.2 语义信息编码模块

Zhang等^[19]提出了SGN,首次使用人体关节的高级语义表示(关节类型与帧间索引),增强了特征表示能力,证明了语义信息的引导相较于单一的关节位置信息在动作识别上效果更加显著。然而SGN原有的语义信息融合方式适用于单层的轻量级网络,却不适用于复杂的多层网络,因为随着网络层数的增加,SGN原有的“关节类型融合—GCN计算—帧间索引融合”先后分块的特征融合方式不能使网络获得更加丰富的信息,反而在深层的网络中语义信息会成为噪声影响识别精度。

本文提出的语义信息编码模块是对SGN原有的特征融合方式进行了改进,将语义信息的融合集成在一个模块内,使其更加适合复杂的多层网络。模块的实现如下:首先,原始的关节位置信息会经过一次编码;然后,通过不同的特征融合方式对不同的语义信息进行融合;最后将融合语义信息后的关节特征信息作为网络输入进行GCN的计算。

1.2.1 关节位置信息编码

在NTU RGB+D等数据集中,骨架关节的位置信息表现为骨架关节的空间三维坐标,即 $\mathbf{J}_{t,k} = (x_{t,k}, y_{t,k}, z_{t,k})^T \in \mathbf{R}^3$ 。本文沿用SGN的编码方式,使用2个全连接层(Fully Connected, FC)对关节位置信息进行编码至更高的维度 C_1 ,具体为

$$\tilde{\mathbf{J}}_{t,k} = \sigma(\mathbf{W}_2(\sigma(\mathbf{W}_1\mathbf{J}_{t,k} + \mathbf{b}_1)) + \mathbf{b}_2) \in \mathbf{R}^{C_1} \quad (1)$$

式中: $\mathbf{W}_1 \in \mathbf{R}^{C_1 \times 3}$, $\mathbf{W}_2 \in \mathbf{R}^{C_1 \times C_1}$ 是2个权值矩阵, $\mathbf{b}_1$ 和 $\mathbf{b}_2$ 表示偏置向量, σ 表示ReLU激活函数。通过编码后的 $\tilde{\mathbf{J}}_{t,k}$ 可以表示更高维度的关节位置信息。

1.2.2 关节类型

将关节独热编码向量 $\mathbf{j}_k \in \mathbf{R}^N$ 代入式(1) 编码计算得到该关节类型的表征向量 $\tilde{\mathbf{j}}_k \in \mathbf{R}^{C_1}$ 。将编码后的关节位置信息 $\tilde{\mathbf{J}}_{t,k}$ 与 $\tilde{\mathbf{j}}_k$ 级联起来得到第 t 帧中第 k 个关节的空间信息表征为

$$\mathbf{z}_{t,k} = [\tilde{\mathbf{J}}_{t,k}, \tilde{\mathbf{j}}_k] \in \mathbf{R}^{2C_1} \quad (2)$$

在第 t 帧中的人体结构所有关节点就可以表示为 $\mathbf{Z}_t = (\mathbf{z}_{t,1}, \mathbf{z}_{t,2}, \dots, \mathbf{z}_{t,N}) \in \mathbf{R}^{N \times 2C_1}$ 。

1.2.3 帧间索引

使用独热编码 $\mathbf{f}_t \in \mathbf{R}^{T_{in}}$ 来表示时间序列总帧数 T_{in} 帧中的某一帧 t 。将 $\mathbf{f}_t$ 代入式(1) 编码得到该帧的帧间索引 $\tilde{\mathbf{f}}_t$, 将 $\tilde{\mathbf{f}}_t$ 与 $\mathbf{z}_{t,k}$ 相加, 这样就能把 $\tilde{\mathbf{f}}_t$ 融合到第 t 帧内某一关节空间信息 $\mathbf{z}_{t,k}$ 中得到关节的时空信息表征:

$$\mathbf{x}_{t,k} = \mathbf{z}_{t,k} + \tilde{\mathbf{f}}_t \in \mathbf{R}^{C_{in}} \quad (3)$$

式中: $C_{in} = 2C_1$, 该帧内所有关节共享一个帧间索引。于是拓展到时间序列中, 原始的人体骨架关节的时空特征信息融合语义信息后可以重新定义为 $\mathbf{X}_{in} \in \mathbf{R}^{C_{in} \times T_{in} \times N}$ 。

对于多层的网络, 本方法使用含卷积块的残差连接^[25]将关节特征信息 $\mathbf{X}_{in}$ 传递到更深层的网络, 这种方式可以避免语义信息在深层的网络中成为噪声而影响动作识别精度。于是, 在语义信息编码模块的语义引导下, 网络可以获取的关节信息更加丰富: 对于一个人体骨架序列, 网络可以通过关节的空间语义(节点类型)区分识别不同关节, 并通过关节的时间语义(帧间索引)区分关节动作先后过程。

1.3 自适应拓扑推理图卷积网络

将经过语义信息编码模块的关节特征信息作为输入,进行GCN的计算。本文的自适应拓扑推理图卷积网络是在2s-AGCN^[18]基础上应用拓扑结构推理网络进一步改进得到的。

1.3.1 拓扑结构推理网络

由于CNN具有高效的特征学习能力,在自适应学习人体关节的拓扑信息提升GCN网络的识别精度上有很大的潜力。然而大多数GCN算法都忽视了这一点,只在GCN的计算方式上或者网络结构上进行改进,使用预定义的邻接矩阵来静态地表示人体拓扑结构,忽略了运动过程中人体不相邻的关节点之间也存在潜在的相关性。

因此本文提出了一种拓扑结构推理网络动态地根据不同的样本学习不同的拓扑结构,如图2所示。

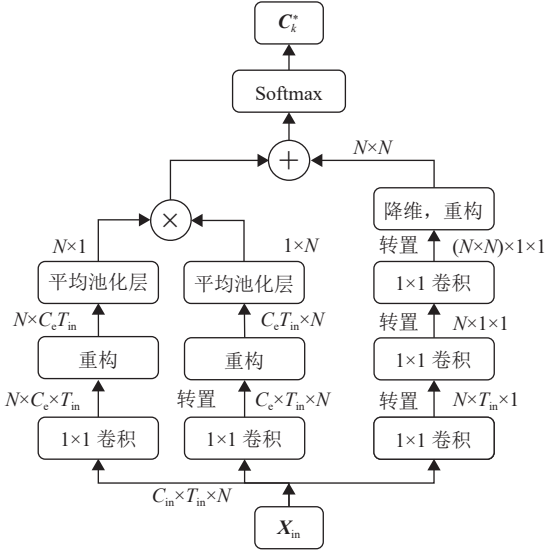


图2 拓扑结构推理网络

Fig.2 Skeleton topology inference network

拓扑结构推理网络由2个部分组成,第1部分是由左右2个对称的分支组成。首先关节的特征信息 X_{in} 作为网络的输入,依次经过2个分支的 1×1 卷积层、张量的重构(reshape)操作、平均池化层来改变维度,2个分支的结果进行矩阵相乘。

拓扑结构推理网络的第2部分充分利用CNN特征学习能力,获取关节特征维度全局的上下文特征信息。将 X_{in} 视为“伪图”,在每次进行 1×1 卷积前会通过张量的转置操作依次将关节特征的维度(位置特征 C_{in} 、时间 T_{in} 、关节点数 N) 变换为“伪图”的通道,进行全局计算。卷积操作完成后进行张量的降维(flatten)操作和重构维度。推理网络的2个部分计算结果相加后,经过一个Softmax层归一化处理得到了

最终学习到的拓扑结构表征 C_k^* ,作为邻接矩阵的一部分参与GCN的计算。

本文的拓扑结构推理网络以全局的方式动态地学习人体骨架关节的拓扑结构。本方法属于自适应图学习的方法,在梯度随机优化过程中不断调整邻接矩阵,学习的过程完全由数据驱动而不需要任何先验知识,可以通过不同类别的动作样本学习不同的图结构。相比于基于先验知识的静态的邻接矩阵,拓扑结构推理网络学习到的邻接矩阵可以反映任意骨架关节点之间的相关性和相关性强弱。本文的拓扑结构推理网络可以应用到一般的基于GCN的算法中,从而提升动作识别的准确率。

1.3.2 改进的自适应图卷积层

本文基于2s-AGCN,改进了原有的自适应图卷积层,如图3所示。

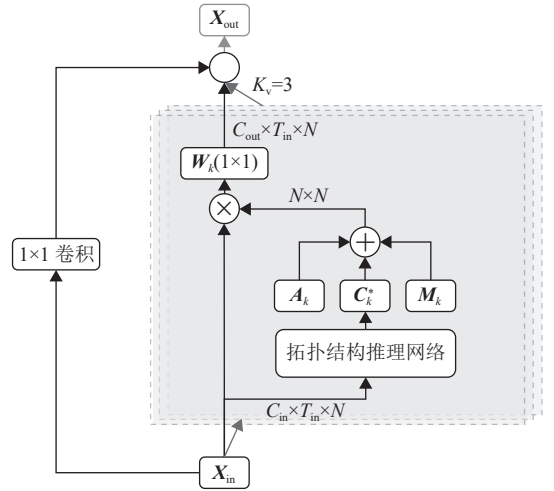


图3 自适应图卷积层

Fig.3 Adaptive graph convolution layer

使用本文提出的拓扑结构推理网络自适应学习得到拓扑结构 C_k^* ,代替原有的参数 C_k 进行GCN的计算。自适应图卷积层计算公式为

$$X_{out} = \sum_k^{K_v} W_k X_{in} (A_k + M_k + C_k^*) \quad (4)$$

式中: X_{in} 为自适应图卷积层的输入,即语义信息编码模块的输出结果, X_{out} 为自适应图卷积层输出结果; K_v 为按照STGCN^[17]的空间划分策略对人体关节划分的子集数量,本文取 $K_v = 3$; W_k 为第 k 个子集的权重矩阵。

GCN的邻接矩阵由 A_k 、 M_k 和 C_k^* 组成。

$$A_k = D_k^{-\frac{1}{2}} \bar{A}_k D_k^{-\frac{1}{2}} \in \mathbf{R}^{N \times N} \quad (5)$$

式中: $\bar{A}_k = (A + I)$ 为原始邻接矩阵与单位矩阵之和,表示人体骨架关节点之间的物理连接关系。计算上

和PR-GCN^[22]类似。 D_k 为度矩阵,是规范化的对角矩阵,其元素 $D_k^{ii} = \sum_j A_k^{ij} + \alpha$ 。为了避免空行, α 通常设置为一个极小值0.001。 $M_k \in \mathbf{R}^{N \times N}$ 类似于STGCN^[17]中的mask M矩阵,发挥图注意力机制的作用,初始化为0,可以学习到任意关节点之间的共性连接方式。 C_k 为拓扑结构推理网络的输出结果。最后通过一个带卷积块的残差连接将自适应图卷积层首尾相连。

1.4 多流自适应拓扑推理图卷积网络

Shi等^[26]证明应用运动信息作为输入的多流网络效果比单流和双流网络更好,因为不同流之间可以形成互补关系。基于该结论,本文除了2s-AGCN原有的人体关节点与骨架信息,还应用了二者的运动信息分别输入到各自网络分支中,共同组成多流图卷积网络。运动信息定义如下。

在一段时序中的某一帧 t 内某一关节点 k 位置信息的三维坐标为 $J_{t,k} = (x_{t,k}, y_{t,k}, z_{t,k})$,下一帧该关节点的坐标为 $J_{t+1,k}$,那么其速度矢量即关节点运动信息可以表示为 $JM_{t,k} = J_{t+1,k} - J_{t,k}$ 。同理,骨架运动信息可以表示为 $BM_{t,k} = B_{t+1,k} - B_{t,k}$ 。特别地,多流网络的4个分支的原始输入数据都会经过语义信息编码模块,得到融合了语义信息的4类不同类型的特征信息,然后同时使用4个相同的自适应拓扑推理图卷积网络进行人体动作特征信息的提取。4个分支的输入信息侧重点不同,可以实现优势互补。

语义引导下多流自适应拓扑推理图卷积网络整体结构如图1所示,由4个分支组成。网络每个分支的末端有1个全连接层和1个Softmax分类器。对4个分支的Softmax分数进行加权求和,Softmax分类器选用交叉熵函数作为网络动作识别的损失函数。

2 实验

2.1 数据集

NTU RGB+D: NTU系列数据集是基于骨架信息的人体动作识别任务使用最为广泛的数据集。NTU RGB+D包含了60类动作的56 880个动作片段,这些片段提供了人体运动过程中25个关节点每个关节点的位置信息(x, y, z)。该数据集有2个基准:(1) Cross-subject (X-Sub) 将一部分测试者的40 320个动作片段用于训练网络,其余的测试者的动作片段用于验证;(2) Cross-view(X-View) 通过2个相机拍摄的37 920个片段用于训练网络,其余的片段由另外的1个相机拍摄用于验证。

NTU RGB+D 120: 该数据集是NTU RGB+D的

拓展版本,包含了120个动作类别,有更多的视频片段。该数据集有2个基准:(1) Cross-subject(X-Sub)与NTU RGB+D一样的设置,数据集中用于训练和验证的骨架信息来自不同的测试者;(2) Cross-setup(X-Setup)包含54 471个片段用于训练,59 477个片段用于验证。

2.2 实验过程

本文的实验均在Ubuntu 20.04操作系统中部署,使用Python语言、Pytorch深度学习框架编写和训练网络。硬件上,使用NVIDIA GeForce RTX 3060显卡进行GPU加速训练。使用随机梯度下降法(Stochastic Gradient Descent, SGD)作为训练优化器。批处理大小Batch size设置为64。网络一共训练65轮次,学习率初始值设置为0.1,在第35、55个轮次分别衰减为原来的0.1倍。

2.3 实验结果分析

2.3.1 消融实验

在本节中,在NTU RGB+D数据集的X-View基准上对本文提出的模型进行消融实验,以动作识别的准确率作为评价指标,使用语义引导下多流自适应拓扑推理图网络中的关节点信息单一分支(单流)验证网络各个部分对人体动作识别的有效性。

语义信息编码模块:如1.2节所述,本文在进行GCN计算的数据前处理时,会在一个语义信息编码模块中融合人体骨架关节点的语义信息,即关节点类型 $\tilde{j}_k$ 和帧间索引 $\tilde{f}_t$ 。在进行消融实验时通过网络融合不同的语义信息来验证语义信息有效性,实验结果如表1所示。

表 1 添加不同类型语义信息的对比结果

添加关节点类型	添加帧间索引	识别准确率/%
×	×	93.9
√	×	94.4
×	√	94.2
√	√	94.7

通过实验结果可以得知,未融合语义信息而仅用单一的关节点位置信息作为输入网络的动作识别准确率仅有93.9%。融合了帧间索引后的输入信息能使识别准确率提升0.3%,这是因为帧间索引作为关节点的时间语义信息,可以引导网络区分关节点动作发生的先后过程,进而区分相同节点的不同动作,如“起立”与“坐下”。相较于帧间索引,关节点类型的空间语义信息能为网络识别准确率带来0.5%的提

升,这是因为在进行基于骨架关节点的动作识别时,GCN会更加关注不同节点空间上的信息,该语义信息会引导网络区分识别不同类型的关节,从而区分不同的动作。

实验还将SGN原有的融合方式应用在本网络中,动作识别准确率只有93.7%,比不含语义信息编码模块的网络效果更差,这是因为SGN网络先进行GCN计算再进行特征融合的处理方式不能使网络获得更加丰富的信息,反而在深层的网络中语义信息会成为噪声影响识别精度。而本文的语义信息编码模块里改进的融合方式成功地将语义信息用于复杂的多层网络,能让网络获取更加丰富的特征信息,从而使网络能够通过节点类型和帧间索引区分细微差别的不同类动作,使动作识别准确率提高了0.8%。在基于骨架关节点的动作识别任务中,本文的语义信息编码模块作为新型的数据处理的方式可以应用于任何基于GCN算法的大型网络,提高识别精度。

自适应图卷积层:消融实验通过使用语义引导下多流自适应拓扑推理图网络中的关节点信息单分支(单流)验证GCN计算时式(4)中邻接矩阵各组成部分的有效性,结果如表2所示。可以看出,如果仅使用固定的邻接矩阵 A_k ,动作识别准确率是最低的,只有93.5%。这是因为所有的样本都共享一个预定义的拓扑结构,也就是一个固定的由人体关节点的物理连接而定义的邻接矩阵,不够灵活,鲁棒性差。原有的2s-AGCN的邻接矩阵组合为 $(A_k + M_k + C_k)$,可以自适应地学习样本中的拓扑结构,有效地改善了这一缺陷,识别精度达到了94.4%。基于自适应学习的思想,本文用拓扑结构推理网络的计算结果 C_k^* 代替2s-AGCN原有的 C_k ,如式(4)所示,得到更高的准确率94.7%。这得益于CNN高效的特征学习能力,拓扑结构推理网络可以通过获取关节点特征维度全局的上下文特征信息,自适应地根据不同的样本学习不同的图结构。这种方式学习到的邻接矩阵摆脱了固定人体拓扑结构的局限,能够有效地表示没有物理直

表 2 自适应图卷积层不同部分的对比结果
Table 2 Comparison results of different parts of adaptive graph convolutional layer

邻接矩阵				识别准确率/%
A_k	M_k	C_k	C_k^*	
√	×	×	×	93.5
√	√	×	×	94.1
√	√	√	×	94.4
√	√	×	√	94.7

接连接的两两关节点的相关性以及相关性的强弱。因为不同的动作关节点间的相关性强弱是不同的,因此本文提出的拓扑结构推理网络有利于动作识别,尤其是复杂的难度动作的识别任务。

本文提出的拓扑结构推理网络同样适用于大多数基于GCN算法的网络,在增加较少参数量的同时能为网络的动作识别精度带来显著的提升。

2.3.2 与其他算法对比

在数据集NTU RGB+D和NTU RGB+D 120上,本文提出的语义引导下的多流自适应拓扑推理图卷积网络与其他人体动作识别算法,包括基于CNN、RNN和GCN算法的对比结果如表3、表4所示。

表 3 NTU RGB+D数据集对比结果
Table 3 Comparison results on NTU RGB+D

方法		识别准确率/%	
		X-Sub	X-View
基于CNN	CNN+Motion+Trans ^[10]	83.2	88.8
	D-Pose Traversal Conv ^[11]	76.8	84.3
	RotClips+MTCNN ^[12]	81.1	87.4
基于RNN	GCA-LSTM ^[13]	77.1	85.1
	SR-TSL ^[14]	84.8	92.4
	dense-IndRNN-aug ^[15]	86.7	93.9
基于GCN	ST-GCN ^[17]	81.5	88.3
	2s-AGCN ^[18]	88.5	95.1
	SGN ^[19]	89.0	94.5
	Sem-GCN ^[24]	86.2	94.2
	3s RA-GCN ^[20]	87.3	93.6
	PGCN-TCA ^[23]	88.0	93.6
	PR-GCN ^[22]	85.2	91.7
	Ours (单流)	89.2	94.7
	Ours (双流)	89.2	95.5
	Ours (多流)	89.3	95.8

表 4 NTU RGB+D 120数据集对比结果
Table 4 Comparison results on NTU RGB+D 120

方法		识别准确率/%	
		X-Sub	X-Setup
基于CNN	RotClips+MTCNN ^[12]	62.2	61.8
基于RNN	GCA-LSTM ^[13]	61.2	63.3
基于GCN	2s-AGCN ^[18]	82.9	84.9
	SGN ^[19]	79.2	81.5
	3s RA-GCN ^[20]	81.1	82.7
	Ours (多流)	83.9	85.8

可以看出在人体动作识别的任务上,由于GCN在处理拓扑信息上有天然的优势,基于GCN的算法性能表现是优于基于CNN和RNN的。本文改进了2s-AGCN,嵌入了语义信息编码模块与拓扑结构

推理网络,集成了多流网络,比单一关节点分支(单流)网络的动作识别准确率提升了1.1%。2s-AGCN网络就是双流的网络,本文将语义信息编码模块与拓扑结构推理网络应用到双流网络上与2s-AGCN进行对比,在NTU RGB+D数据集的Cross-subject基准上准确率提升了0.7%,在Cross-view基准上准确率提升了0.4%,证明了本文提出的语义信息编码模块和拓扑结构推理网络可以有效提升GCN识别精度。本文应用了运动流,集成了语义引导下多流自适应拓扑推理图卷积网络,在NTU RGB+D与NTU RGB+D 120数据集上的最终实验结果均达到了目前动作识别算法的先进水平。

3 结语

本文改进了SGN原有的语义信息(关节点类型和帧间索引)融合方式,集成到一个语义信息编码模块,使语义信息能够成功应用于大型的多层网络,在语义引导下为网络提供了更加丰富的特征信息,从而提高算法的动作识别准确率。此外,在进行GCN的计算时,本文提出了一个拓扑结构推理网络,结合CNN高效的特征学习能力,能够自适应地根据不同样本的上下文特征信息学习到不同的拓扑结构,这种方式学习到的邻接矩阵摆脱了人体固定拓扑结构的局限。本文将语义信息编码模块和拓扑结构推理网络嵌入2s-AGCN网络中,并增加了相关运动流分支,提出一种语义引导下多流自适应拓扑推理图卷积网络。通过对比实验证明,本文的算法达到了目前动作识别算法的先进水平。另外,本文的语义信息编码模块和拓扑结构推理网络可以用于一般的基于GCN的大型多层网络,有效提高网络的识别精度,为基于骨架关节点信息的人体动作识别算法提供新的思路。本文只关注拓扑结构上的自适应学习来提升网络识别精度,未来将尝试使用更加高效的注意力机制对通道或者其他维度上的特征进行动态权重调整,进一步提升网络的性能。

参考文献:

- [1] EVANGELIDIS G, SINGH G, HORAUD R. Skeletal quads: human action recognition using joint quadruples[C]//2014 22nd International Conference on Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2014: 4513-4518.
- [2] 周小平, 郭开仲. 基于计算机视觉的腾空飞脚错误动作识别模型[J]. 广东工业大学学报, 2012, 29(4): 14-17.
ZHOU X P, GUO K Z. The model for the recognition of flying kick error action based on computer vision [J]. Journal of Guangdong University of Technology, 2012, 29(4): 14-17.
- [3] WANG H, SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2013: 3551-3558.
- [4] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.
- [5] SADEGH A M, SADAT S F, SALZMANN M, *et al.* Encouraging lstrms to anticipate actions very early[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 280-289.
- [6] JAIN A, SINGH A, KOPPULA H S, *et al.* Recurrent neural networks for driver activity anticipation via sensory-fusion architecture[C]//2016 IEEE International Conference on Robotics and Automation (ICRA) . New York: IEEE, 2016: 3118-3125.
- [7] WANG H, WANG L. Beyond joints: learning representations from primitive geometries for skeleton-based action recognition and detection [J]. IEEE Transactions on Image Processing, 2018, 27(9): 4382-4394.
- [8] YANG C, XU Y, SHI J, *et al.* Temporal pyramid network for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2020: 591-600.
- [9] WANG L, TONG Z, JI B, *et al.* TDN: temporal difference networks for efficient action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2021: 1895-1904.
- [10] 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) . New York: IEEE, 2017: 597-600.
- [11] WENG J, LIU M, JIANG X, *et al.* Deformable pose traversal convolution for 3D action and gesture recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV) . Cham, Switzerland: Springer, 2018: 136-152.
- [12] KE Q, BENAMOUN M, AN S, *et al.* Learning clip representations for skeleton-based 3d action recognition [J]. IEEE Transactions on Image Processing, 2018, 27(6): 2842-2855.
- [13] LIU J, WANG G, DUAN L Y, *et al.* Skeleton-based human action recognition with global context-aware attention LSTM networks [J]. IEEE Transactions on Image Processing, 2017, 27(4): 1586-1599.
- [14] SI C, JING Y, WANG W, *et al.* Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]//Proceedings of the European Conference on Computer Vision (ECCV) . Cham, Switzerland: Springer, 2018: 103-118.
- [15] LI S, LI W, COOK C, *et al.* Independently recurrent neural network (indrnn) : building a longer and deeper rnn[C]//Proceedings of the IEEE Conference on Computer Vision and

- Pattern Recognition. New York: IEEE, 2018: 5457-5466.
- [16] CHEN T, ZHOU D, WANG J, *et al.* Learning multi-granular spatio-temporal graph network for skeleton-based action recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2021: 4334-4342.
- [17] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2018, 32(1).
- [18] SHI L, ZHANG Y, 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. New York: IEEE, 2019: 12026-12035.
- [19] ZHANG P, LAN C, ZENG W, *et al.* Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 1112-1121.
- [20] SONG Y F, ZHANG Z, SHAN C, *et al.* Richly activated graph convolutional network for robust skeleton-based action recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(5): 1915-1925.
- [21] ZENG A, SUN X, YANG L, *et al.* Learning skeletal graph neural networks for hard 3D pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 11436-11445.
- [22] LI S, YI J, FARHA Y A, *et al.* Pose refinement graph convolutional network for skeleton-based action recognition [J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1028-1035.
- [23] YANG H, GU Y, ZHU J, *et al.* PGCN-TCA: pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition [J]. IEEE Access, 2020, 8: 10040-10047.
- [24] DING X, YANG K, CHEN W. A semantics-guided graph convolutional network for skeleton-based action recognition[C]//Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence. New York: Association for Computing Machinery, 2020: 130-136.
- [25] HE K, ZHANG X, REN S, *et al.* Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2016: 770-778.
- [26] 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.

(责任编辑: 杨耀辉)