计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900059

基于动态拓扑图的人体骨架动作识别算法

解　宇[1]　杨瑞玲[1]　刘公绪[2]　李德玉[1]　王文剑[1]
1 山西大学计算机与信息技术学院　太原 030006
2 西安电子科技大学电子工程学院　西安 710071
(yuxie@sxu.edu.cn)

摘　要　传统的人体骨架动作识别算法采用手动构建拓扑图的方式来建模包含在多个视频帧中的动作序列,并针对性地学习每个视频帧以反映数据变化,这容易造成计算代价大、网络泛化性低和灾难性遗忘等问题。针对上述问题,提出了基于动态拓扑图的人体骨架动作识别算法,使用持续学习思想动态构建人体骨架拓扑图。将具有多关系特性的人体骨架序列数据重新编码为关系三元组,并基于长短期记忆网络,通过解耦合的方式学习特征嵌入。当处理新骨架关系三元组时,使用部分更新机制动态构建人体骨架拓扑图,并采用基于时空图卷积网络的骨架动作识别算法来实现动作识别。实验结果表明,所提方法在Kinetics-Skeleton,NTU-RGB+D(X-Sub)和NTU-RGB+D(X-View)基准数据集上分别取得了40%,85%和90%的识别准确率,提高了人体骨架动作识别的准确率。

关键词:人体动作识别;人体骨架数据;灾难性遗忘;持续学习;图卷积网络
中图法分类号　TP391

Human Skeleton Action Recognition Algorithm Based on Dynamic Topological Graph

XIE Yu[1],YANG Rui-ling[1],LIU Gong-xu[2],LI De-yu[1] and WANG Wen-jian[1]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 School of Electronic Engineering,Xidian University,Xi'an 710071,China

Abstract　Traditional human skeleton action recognition algorithms manually construct topological graphs to model the action sequence contained in multiple video frames and learn each video frame to reflect the data changes,which may lead to the high computational cost,low network generalization performance and catastrophic forgetting. To solve these problems,a human skeleton action recognition algorithm based on dynamic topological graph is proposed,in which the human skeleton topological graph is dynamically constructed based on continuous learning. Specifically,human skeleton sequence data with multi-relationship characteristics are recoded into relationship triplets,and feature embedding is learned in a decoupling manner via the long short-term memory network. When handling new skeleton relationship triplets,we dynamically construct the human skeleton topological graph by a partial update mechanism,and then send it to the skeleton action recognition algorithm based on spatio-temporal graph convolution network for action recognition. Experimental results demonstrate that the proposed algorithm achieves 40%,85% and 90% recognition accuracy on three benchmark datasets,namely Kinetics-Skeleton,NTU-RGB+D(X-Sub) and NTU-RGB+D(X-View),respectively,which improve the accuracy of human skeleton action recognition.

Keywords　Human action recognition,Human skeleton data,Catastrophic forgetting,Continual learning,Graph convolution network

1　引言

人体动作识别是计算机视觉领域的重要任务,现有的人体动作识别方法主要分为两个分支[1]:1)基于视频的动作识别,从人体外观、人体深度数据和光流等角度实现动作识别;2)基于人体骨架数据的动作识别,从骨架运动轨迹中提取动作特征以实现动作识别。人体骨架序列数据中包含更为准确的关节位置信息,使基于人体骨架数据的动作识别方法比

到稿日期:2021-09-07　返修日期:2021-09-22
基金项目:国家自然科学基金(62076154,62106131,62106134);中央引导地方科技发展资金项目(YDZX20201400001224);山西省国际科技合作计划项目(201903D421050)
This work was supported by the National Natural Science Foundation of China(62076154,62106131,62106134),Program of Central Funds Guiding the Local Science and Technology Development(YDZX20201400001224) and Key R&D program of Shanxi Province(International Cooperation)(201903D421050).
通信作者:王文剑(wjwang@sxu.edu.cn)

基于视频的动作识别方法在处理视点变换、身体比例和运动速度等问题时均具有更好的鲁棒性。

人体骨架动作识别的目标是识别人体骨架构成的时间序列代表的动作类型。传统的人体骨架动作识别方法主要分为基于手工特征的方法和基于深度学习的方法。基于手工特征的方法通过人工设计特征的方式来捕捉关节之间的动力学关系[2]。基于深度学习的方法采用递归神经网络等深层模型,以端到端的方式建模人体各部分的关节点[3-7]。后者主要包含3种框架:基于序列的方法使用神经网络对按一定规则排列的关节序列建模;基于图像的方法使用卷积神经网络对编码为伪图像的人体骨架序列建模;基于图的方法以人体关节点为顶点,以关节点之间的自然连接为边,为关节数据构建拓扑图,同时将骨架数据的高维特征作为辅助输入[5-7],然后使用图卷积神经网络对人体骨架数据建模。

基于图的方法手动为关节数据构建拓扑图并辅以高维特征,然而实际场景中的一个动作通常由多个视频帧构成,手动构建拓扑图的方式难以反映关节点、骨骼以及连接关节点和未连接关节点之间的依赖关系。对于一个动作包含的每个视频帧,均需针对性地训练模型并不断更新参数以反映数据变化,这将导致计算代价大且容易出现灾难性遗忘问题。

为解决上述问题,本文提出了基于动态拓扑图的人体骨架动作识别算法,使用持续学习思想,基于长短期记忆网络(Long Short-Term Memory,LSTM)动态构建人体骨架拓扑图,并基于建模后的人体骨架实现动作识别。具体地,使用关节类型、关节时间位移、关节空间位移和帧编号将人体骨架数据编码为关系三元组,节点间的关系为关节帧所属的动作类型,然后采用解耦合的方式学习三元组数据的图嵌入。当新的关系数据到来时,使用部分更新机制更新与新骨架关系三元组有关的旧骨架关系三元组,动态构建人体骨架拓扑图,最后采用基于时空图卷积网络的骨架动作识别算法来实现动作识别。

本文工作的主要贡献在于:

(1)提出了基于动态拓扑图的人体骨架动作识别算法,通过解耦合的方式动态构建骨架拓扑图,并在训练过程中引入评估机制,以保证网络能准确识别已学到的动作类别,解决训练过程中的灾难性遗忘问题。

(2)将有多关系特性的人体骨架数据编码为关系三元组,采用持续学习思想,使网络按动作类别学习序列帧实体的解耦合特征,并使用部分更新策略构建动态更新的人体骨架拓扑图,提升网络的泛化性。

(3)公开数据集上的实验结果表明,本文算法能有效提升动作识别任务的准确率。

2　相关工作

2.1　基于图卷积的人体骨架动作识别算法

针对人体动作识别任务,文献[3]提出了基于时空图卷积网络的人体骨架动作识别算法(Spatial-Temporal Graph Convolutional Networks,ST-GCN),首次将图卷积应用于人体骨架动作识别领域,通过空间图卷积提取并聚合单个视频帧中人体骨架序列的空间特征,利用时间图卷积提取并聚合同一关节点在不同视频帧中的运动特征,使人体骨架特征图同时包含时间运动特征和空间运动特征。此外,文献[3]设计了子集分隔策略,依据运动特征划分目标节点邻域,使网络在识别不同动作特征时关注不同的关节点集,以进一步提升网络对于不同肢体动作的学习能力。以该算法为基础,许多基于图卷积的人体骨架动作识别算法被提出。

基于图卷积的人体骨架动作识别算法可分为时空特征提取、特征融合和动作分类3部分,首先使用图卷积网络,从建模为骨架拓扑图的人体骨架数据中提取时空特征,然后融合学习到的特征,并基于融合后的特征实现动作分类。时空特征提取部分旨在学习骨架拓扑图中节点的特征表示,主要包括以下4类方法:1)高维特征提取方法计算骨架高维特征作为辅助特征输入,以丰富关节特征,如Ding等[7]计算关节点权重和姿态权重,Tian等[8]计算关节点时序散度;2)特征构建方法通过向拓扑图中添加特征,来增加关节点邻域中的信息,如Shi等[9]添加了骨骼方向,Tang等[10]人为构建了部分关节点连接;3)骨骼划分策略依据人体骨骼结构将骨架划分为多个部分,如Thakkar等[11]为每个肢体部分中的关节点定义了特征提取方式;4)拓扑图重构方法拓展了图结构中顶点和边的概念,如Li等[12]通过将人体骨骼抽象为图节点来构建部件拓扑图。特征融合部分旨在将特征提取得到的节点特征聚合为高维特征,然后输入动作分类部分实现动作分类,如Li等[13]使用拼接操作融合学习时间特征图和空间特征图,并将融合后的特征图输入ST-GCN网络以实现动作分类。

以上方法通过手工构建关节特征图来进行特征提取的方式,难以学习到部分关键关节位置以及运动信息,同时忽略了在真实场景中会不断有新数据送入网络,导致模型每次都需重新训练并更新参数以学习新数据特征,计算代价较大且容易出现灾难性遗忘问题。

2.2　持续学习

神经网络通过不断更新参数的过程从任务中学习新知识,但在学习新任务时存储新知识的参数会覆盖掉存储旧知识的参数,从而导致灾难性遗忘问题。持续学习(continual learning)[14]主要用于解决两个问题:1)提升模型可塑性,即如何利用从之前任务中学到的知识使网络更好地学习新任务;2)提升模型稳定性,即在学习新任务时如何保留从之前任务中学到的知识。传统的持续学习思想包括以下两种:1)基于Regularization的方法[15]为旧数据赋予重要性参数,并在学习新数据时不断调整;2)基于Memory的方法[16-17]记忆一些旧数据,与新数据共同学习。Liu等[18]使用LSTM有效地建模序列化数据中的时序信息。为了解决语义问答任务建模中的灾难性遗忘问题,Kou等[19]针对该任务中的多关系数据流,提出了基于解耦合的框架,有效提高了语义问答任务的准确率。人体骨架数据具有多关系特性,因此可使用持续学习的方法来解决动作识别任务在模型训练过程中的灾难性遗忘问题,但目前仍没有较好的持续学习方案可用于实现动态构建人体骨架拓扑图。

3 基于动态拓扑图的人体骨架动作识别算法

3.1 基本定义

定义 1(人体骨架序列) 人体骨架序列通常由连续的人体骨架帧构成,每个人体骨架帧是一个由一系列关节点的 2D 或 3D 坐标以及坐标的置信度所构成的集合。定义人体骨架序列 $VT=\{VT_1,VT_2,\cdots,VT_T\}$,其中 T 为人体骨架帧的数量,VT_i 为人体骨架序列中第 i 个人体骨架帧。VT_i 由一系列包含时空信息的关节点组成,即 $VT_i=\{V_1,V_2,\cdots,V_N\}(1\leqslant i\leqslant T)$,其中 N 为人体骨架帧中关节点的数目。定义人体骨架帧的第 i 个关节点 $V_i(1\leqslant i\leqslant N)=(x_i,y_i,score_i)$,其中 (x_i,y_i) 为该关节点的位置信息,$score_i$ 为该关节点位置的置信度信息。人体骨架序列的维度为 (T,N,C),其中,T 为人体骨架帧的数量,N 为关节点数目,C 为位置向量维度(V_i 的位置向量维度为 3)。

定义 2(时空图|多关系图) 为了便于使用图卷积提取人体骨架序列的空间特征和时间特征,通常将人体骨架序列重构为时空图 $G=(V,E)$,图中包含关节点集 V 和边集 E。定义关节点集 $V=\{V_{ti}\mid t=1,\cdots,T,i=1,\cdots,N\}$,包括人体骨架序列中的所有关节点。定义边集 $E=\{E_S,E_F\}$,其由帧内人体骨架连接 E_S 和帧间连接 E_F 两个子集构成。定义帧内人体骨架连接 $E_S=\{V_{ti}V_{tj}\mid(i,j)\in H\}$,其中 H 为自然连接的人体关节集合。定义帧间连接 $E_F=\{V_{ti}V_{(t+1)i}\}$,一个特定关节在 E_F 中的所有边可以表示该关节随时间的运动轨迹。

定义 3(关系三元组) 节点之间存在多对多关系的图结构称为多关系图,多关系图可形式化为一组关系三元组 $\{(u,r,v)\}\subseteq V\times E\times V$。其中,$u$ 和 v 为实体,r 为实体之间的某种关系。

3.2 算法描述

为了解决传统的基于图卷积的动作识别算法在学习新数据过程中存在的泛化性不足和灾难性遗忘问题,本文融合持续学习思想,在 ST-GCN 网络的基础上提出了一种基于动态拓扑图的人体骨架动作识别算法。该算法基于人体骨架数据本身的多关系特征,将人体骨架数据重构为人体骨架三元组序列,使用解耦合的方式在多关系数据上学习嵌入特征,并采用部分更新策略动态构建人体骨架序列的关系拓扑图,最终将包含丰富关系的拓扑图输入 ST-GCN 网络以实现动作识别。算法的总体架构如图 1 所示,分为数据预处理、特征解耦、动态拓扑图构建和动作类别预测。

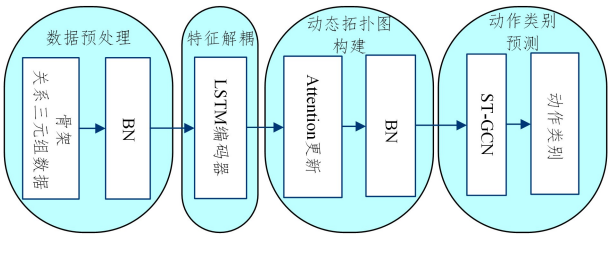

图 1 基于动态拓扑图的人体骨架动作识别算法的结构图

Fig. 1 Structure diagram of human skeleton action recognition algorithm based on dynamic topological graph

3.2.1 数据预处理

数据预处理首先归一化处理多个人体骨架序列 VT 构成的数据集中的关节特征向量,将不同位置的骨架序列转换到相同位置,以促进算法更好地收敛,然后将多个人体骨架序列 VT 构成的数据集重新编码为关系三元组序列构成的数据集。具体地,依据人体骨架数据本身的多关系特征,定义人体骨架关系三元组序列 $RT=\{(VT_i,r_{ij},VT_j)\mid 1\leqslant i\leqslant T,1\leqslant j\leqslant T\}$,其中,$VT_i$ 为人体骨架序列集合中第 i 帧关节点的集合,VT_j 为人体骨架序列集合中第 j 帧关节点的集合,r_{ij} 为该人体骨架序列集合所表示的动作类别。

3.2.2 特征解耦

特征解耦模块从骨架关系三元组数据集中学习实体特征嵌入向量和关系特征嵌入向量。基于不同动作中各独立关节点之间的关联关系,使用 LSTM 网络将骨架关系三元组序列 RT 中的视频帧解耦成多个特征以学习实体特征的嵌入特征向量,同时对动作类别进行编码以学习关系特征的嵌入向量。具体地,对于人体骨架关系三元组序列,实体为单帧视频序列,关系为该视频帧的动作类别。特征解耦通过实体特征解耦和关系特征解耦,将骨架关系三元组解耦成多个特征。实体特征解耦将单个视频帧依据人体骨架序列中的关节点类型解耦为多个特征,然后分别学习每个关节特征的嵌入特征向量,即将每个单视频帧表示的实体 $VT_i\in V$ 转化为由多个独立关节点分量构成的集合 $VT_L=[V_1,V_2,\cdots,V_N]$,其中 $VT_L\in R^d$,d 为向量维数。关系特征解耦将动作类型进行 one-hot 编码,并使用该编码表示动作类别的嵌入特征向量。

3.2.3 动态拓扑图构建

动态拓扑图构建模块采用持续学习思想,基于特征解耦得到的特征向量,为各个类型的动作构建骨架拓扑图。具体地,使用 K-means 算法对特征解耦输出的表示动作类别的关系特征嵌入向量进行聚类,并依据聚类中心将数据集划分为多个训练集。然后按 one-hot 编码后的动作类别顺序,采用 attention 机制和部分更新策略构建不同类型动作的拓扑图。同时,在每一批次数据训练结束后,使用动作识别准确率来评估模型性能,以保证模型在学习新数据后仍能在已学习过的数据集上取得较好的评估结果。

定义关系三元组序列 $\mathcal{T}_i=\{(u_1^{\mathcal{T}_i},r_1^{\mathcal{T}_i},v_1^{\mathcal{T}_i}),\cdots,(u_m^{\mathcal{T}_i},r_m^{\mathcal{T}_i},v_m^{\mathcal{T}_i})\}$ 为第 i 个训练集,其中 m 是 $\mathcal{T}_i$ 的实例号。对于一个骨架关系三元组 $(VT_i,r_{ij},VT_j)\in\mathcal{T}_i$,使用 attention 机制提取连续的视频帧 VT_i 及其相对于关系 r_{ij} 的最相关关节点特征,为骨架关系三元组 (VT_i,r_{ij},VT_j) 赋予 N 个注意力权重 $(\alpha_r^1,\alpha_r^2,\cdots,\alpha_r^N)$,其中第 i 个关节点在当前关系 r 中的重要性 α_r^i 定义如下:

$$\alpha_r^i=\frac{\exp(a_r^i)}{\sum_{j=1}^{N}\exp(a_r^j)} \tag{1}$$

其中,a_r^j 为第 j 个关节点在当前关系 r 中的重要性。然后选择 attention 权重最高的视频帧 VT_i 及其前 i 个关节点,利用前 i 个相关关节点中的不同特征构建该动作类别的关系拓扑图。如图 2 所示,单个骨架帧包含 18 个嵌入特征向量,其中粉色的腿部节点为与该动作最相关的前 6 个关节特征。

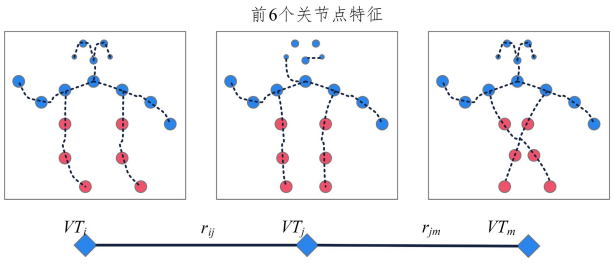

图 2　特征编码示意图(电子版为彩色)

Fig. 2　Schematic diagram of feature coding

为解决模型在学习新数据特征并更新参数的过程中存在的灾难性遗忘问题，采用部分更新策略动态更新拓扑图，即当新的骨架关系三元组到达时，首先从已学习的骨架关系图中激活与新骨架关系三元组具有相关关节点特征的旧骨架关系三元组，并且只更新相关性较高的关节点特征。该过程主要包括以下两个步骤。

(1)邻居激活(neighbor activation)。每出现一个新的数据，网络就预测得出一个新的关系(分类标签)时，更新模块都需要从学习过的三元组序列集合$\{\mathcal{T}_1,\mathcal{T}_2,\cdots,\mathcal{T}_{i-1}\}$中寻找待更新的骨架三元组。由于多关系图中的各节点之间存在多条连接，新出现的数据会对多个旧邻居节点产生影响，因此本文对于每个骨架关系三元组(VT_i,r_{ij},VT_j)，只激活与其直接相连的节点以及距离为 2 的间接相连节点。同时，为减少计算量，本文依据前 i 个关节点的特征相似度，在激活的邻居节点中执行选择更新操作。

(2)部分关节点更新(joints updating)。在更新被激活的邻居骨架关系三元组时，不需要更新已激活邻居的所有关节点特征，只更新相似度最高的公共关节点特征，即更新前 i 个关节点的特征。

3.2.4　动作类别预测

动作类别预测模块将学习到的所有骨架关系三元组数据按帧编号排序，并重构为时空图 G，然后将时空图 G 输入 ST-GCN 网络以实现动作分类预测。

3.3　算法训练

对于新的多关系图数据$\mathcal{T}_i$，在$\mathcal{T}_i$及其激活的相邻骨架关系三元组上使用交叉熵损失函数迭代训练模型。将在新骨架关系三元组序列$\mathcal{T}_i$上迭代训练模型的损失函数表示为$\mathcal{L}_{\text{old}}$，在已激活的旧骨架关系三元组上迭代训练模型的损失函数表示为$\mathcal{L}_{\text{new}}$，$\mathcal{L}_{\text{old}}$和$\mathcal{L}_{\text{new}}$的定义方式相同，$\mathcal{L}_{\text{new}}$的定义如下：

$$\mathcal{L}_{\text{new}}=-\sum_{VT_i\in N(\mathcal{T}_i)}\frac{1}{|c|}\sum_{c=1}^{|c|}y(c)\cdot\log(\sigma(\boldsymbol{w}VT_i)) \tag{2}$$

其中，c 为节点所属动作类别，若节点的动作类别为 c，则 $y(c)$ 为 1，否则 $y(c)$ 为 0；$N(\mathcal{T}_i)$为新骨架关系三元组序列$\mathcal{T}_i$中的节点集；$|c|$为动作类别的数量；$\boldsymbol{w}\in R^{|c|\times d}$为权重矩阵。

最相关关节点特征的数量会影响解耦合效果，因此定义正则化项$\mathcal{L}_{\text{norm}}$约束损失函数，使得最相关关节特征的注意力权重总和接近 1，具体定义如下：

$$\mathcal{L}_{\text{norm}}=\sum_{(VT_L,r,VT_R)\in\mathcal{T}_i}(1-\sum_{i}^{n}\alpha_r^i) \tag{3}$$

其中，n 为所选特征关节点的数目；$VT_L,VT_R\in\boldsymbol{R}^d$和$r\in\boldsymbol{R}^l$为一组人体骨架关系三元组$\mathcal{T}_i$的嵌入特征向量，$d$ 和 l 为向量维数。

定义 β 为正则化项的超参数，算法的总体损失函数 $\mathcal{L}$ 的定义如下：

$$\mathcal{L}=\mathcal{L}_{\text{old}}+\mathcal{L}_{\text{new}}+\beta\cdot\mathcal{L}_{\text{norm}} \tag{4}$$

通过优化该损失函数，来解决传统基于图卷积的动作识别算法在更新参数过程中的灾难性遗忘问题，增强拓扑图中各个节点间所表示的特征，进而提升人体骨架动作识别任务的准确率。

4　实验

4.1　评测数据集

实验采用人体骨架动作识别数据集 Kinetics-Skeleton 和 NTU-RGB+D 来验证本文算法的性能。Kinetics-Skeleton 数据集由 Google DeepMind 通过 OpenPose 姿态估计软件识别到的视频中所有人体骨架关键点信息构成。该数据集将动作分为 400 个类别，共有约 30 万个视频，每帧骨架数据包含 18 个人体关节点位置和置信度得分。NTU-RGB+D 数据集将动作分为 60 个类别，共有约 6 万个视频，包含分辨率为 1920×1080 的 RGB 视频、分辨率为 512×424 的深度图序列、红外视频和包含 25 个人体关节点三维位置的 3D 骨架数据。该数据集采用了多人(X-Sub)与多视角(X-View)两种划分标准，分别按照人物 ID 和采样相机视角划分训练集和测试集。

本实验将 Kinetics-Skeleton 数据集、NTU-RGB+D(X-Sub)数据集和 NTU-RGB+D(X-View)数据集中的分类标签编码为关系三元组中的对应关系，将单个视频帧中人体动作的关节点序列编码为关系三元组中的对应实体，然后将其按动作类别划分为多个训练集，从训练集中为每个动作类别选择一个动作序列作为标准动作序列，以构成标准集。

4.2　实验参数设置

实验中所有算法都基于最优参数进行设置。本文模型主要包含双向 LSTM 编码层和 Soft_max 分类层，使用双向 LSTM 编码层提取特征，然后使用 Soft_max 分类层预测动作类别，用于评估模型效果。实验使用 Adam 优化器训练模型，$r=0.001$。初始化双向 LSTM 编码层的参数值：$max\_length=300$，$input\_size=75$，$hidden\_size=200$。初始化 Softmax 分类层的各项参数：输入向量维度为双向 LSTM 的输出向量维度，类别数为数据集中的动作类别数，不使用偏置项和随机失活；损失函数中正则化项的超参数 β 设为 0.1；其他对比算法都采用各自参数的最佳设置。

4.3　实验结果分析

4.3.1　3 个数据集上的分类准确率分析

本节分别在 Kinetics-Skeleton 数据集、NTU-RGB+D(X-Sub)数据集和 NTU-RGB+D(X-View)数据集上使用动态拓扑图构建模块来动态构建多关系图，在每个类别的数据集中输入模型后，使用 Softmax 层输出动作类别的 Top-1 准确率来评估算法性能。表 1—表 3 列出了 3 个数据集上各个动作类别的 Top-1 准确率。

表 1 本文方法在 Kinetics-Skeleton 数据集上的分类准确率(前 60 个类别)

Table 1 Classification accuracy of our method on Kinetics-Skeleton dataset (top 60 categories)

Class	1	2	3	4	5	6	7	8	9	10
Accuracy	0.79	0.70	0.35	0.50	0.49	0.57	0.98	0.94	0.70	0.48
Class	11	12	13	14	15	16	17	18	19	20
Accuracy	0.45	0.25	0.93	0.95	0.54	0.78	0.54	0.66	0.43	0.55
Class	21	22	23	24	25	26	27	28	29	30
Accuracy	0.46	0.67	0.88	0.87	0.56	0.43	0.23	0.65	0.83	0.77
Class	31	42	33	34	35	36	37	38	39	40
Accuracy	0.67	0.49	0.45	0.56	0.67	0.76	0.45	0.56	0.68	0.78
Class	41	42	43	44	45	46	47	48	49	50
Accuracy	0.76	0.51	0.57	0.69	0.25	0.78	0.48	0.78	0.47	0.98
Class	51	52	53	54	55	56	57	58	59	60
Accuracy	0.79	0.68	0.59	0.87	0.96	0.75	0.66	0.54	0.88	0.65

表 2 本文方法在 NTU-RGB+D(X-Sub)数据集上的分类准确率

Table 2 Classification accuracy of our method on NTU-RGB+D(X-Sub) dataset

Class	1	2	3	4	5	6	7	8	9	10
Accuracy	0.99	1.0	0.49	0.50	0.99	0.27	1.00	0.94	0.80	0.49
Class	11	12	13	14	15	16	17	18	19	20
Accuracy	0.65	0.25	0.43	0.97	0.52	1.00	0.74	0.57	0.67	0.53
Class	21	22	23	24	25	26	27	28	29	30
Accuracy	0.57	0.99	1.00	0.64	0.96	0.99	0.55	0.95	0.69	0.97
Class	31	42	33	34	35	36	37	38	39	40
Accuracy	1.00	0.78	1.00	1.00	0.86	1.00	0.67	0.87	1.00	0.98
Class	41	42	43	44	45	46	47	48	49	50
Accuracy	0.96	0.61	0.55	0.99	0.25	0.98	0.58	0.58	0.97	0.78
Class	51	52	53	54	55	56	57	58	59	60
Accuracy	0.59	0.88	0.99	0.47	1.00	1.00	1.00	1.00	0.33	1.00

表 3 本文方法在 NTU-RGB+D(X-View)数据集上的分类准确率

Table 3 Classification accuracy of our method on NTU-RGB+D(X-View) dataset

Class	1	2	3	4	5	6	7	8	9	10
Accuracy	1.00	0.55	0.99	0.60	0.54	0.67	0.83	0.93	0.80	0.80
Class	11	12	13	14	15	16	17	18	19	20
Accuracy	0.67	0.75	0.88	0.15	0.20	0.99	0.59	0.72	0.54	0.55
Class	21	22	23	24	25	26	27	28	29	30
Accuracy	0.80	0.99	1.00	0.98	0.84	0.59	0.94	0.99	0.96	0.34
Class	31	42	33	34	35	36	37	38	39	40
Accuracy	0.94	0.97	0.98	0.86	0.59	0.66	0.49	0.76	0.89	1.00
Class	41	42	43	44	45	46	47	48	49	50
Accuracy	1.00	1.00	1.00	0.37	0.61	1.000	1.000	1.000	0.39	0.99
Class	51	52	53	54	55	56	57	58	59	60
Accuracy	1.00	0.49	1.00	0.96	1.00	1.00	1.00	0.63	1.00	1.00

在以上 3 个数据集上的实验结果表明,本文提出的基于动态拓扑图的人体骨架动作识别算法能够较为准确地识别出视频序列中的动作类别,且部分类别的识别准确率高达 100%。然而,个别动作的识别准确率较低,如:Kinetics-Skeleton 数据集中 Class45 的识别准确率为 25%;NTU-RGB+D(X-Sub)数据集中 Class59 的识别准确率为 33%;NTU-RGB+D(X-View)数据集中 Class49 的识别准确率为 39%。原因为个别动作的变化幅度过小、关节位置变化不明显,使得网络从中提取的关节特征变化不明显,从而导致网络对该类别动作的识别准确率较低。此外,本文算法在 Kinetics-Skeleton 数据集上的识别准确率较低,原因是 2D 视频中存在关节点遮挡问题,导致基于 OpenPose 姿态识别得到的 Kinetics-Skeleton 数据集中关节位置的准确率低。但从总体结果来看,本文算法具有较高的骨架动作识别准确率。

4.3.2 整体性能和平均性能评估

本节从整体性能(whole performance)和平均性能(average performance)两个角度评估基于动态拓扑图的人体骨架动作识别算法的性能。评估整体性能时,将所有数据集一次性送入网络,并计算对整个测试集的识别准确率。评估平均性能时,将测试集按类别送入网络,并在训练过程中执行评估,最后计算所有测试集评估指标的平均值。由于平均性能能够突出算法在处理灾难性遗忘问题时的性能,因此它是评估模型的主要指标。

由图 3 可知,基于动态拓扑图的人体骨架动作识别算法在 3 个数据集上的平均性能均优于整体性能,这充分说明本文提出的动态拓扑图构建策略将数据集按类别输入网络的方式能够使网络更为集中地学习特定动作类别的特征,从而减小其他类别动作对当前识别任务的影响,解决了网络更新参数过程中的灾难性遗忘问题,提升了网络的泛化性。此外,本文算法在单帧人体骨架序列包含 25 个关节点的 NTU-RGB+D(X-View)数据集上的平均性能达到了 90%,显著优于单帧人体骨架序列包含 24 个关节点的 NTU-RGB+D(X-Sub)

数据集的平均准确率(85%)，表明识别性能与特征维数呈正相关。

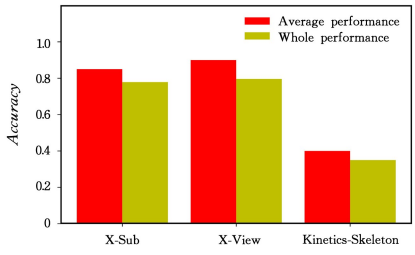

图 3　整体性能和平均性能的对比图

Fig. 3　Comparison between whole performance and average performance

4.3.3　与不同基线算法的性能对比

本节以 Top-1 准确率为评价指标，将基于动态拓扑图的人体骨架动作识别算法的性能与 ST-GCN 算法、基于神经搜索学习图卷积网络的人体动作识别算法(GCN-NAS)[20]和基于时空转换器的人体骨架动作识别算法(ST-TR)[21]的性能进行对比。图 4 中的实验结果表明，相比 3 种基线算法，本文算法使用动态构建拓扑图的方式替代手动构建动态拓扑图的方式增加了特征维数，使得本文算法在 Kinetics-Skeleton 数据集上的准确率提升至 40%，在 NTU-RGB+D(X-Sub)数据集和 NTU-RGB+D(X-View)数据集上的准确率分别达到了 85%和 90%，提升了基于骨架的人体动作识别算法的准确率。本文动态构建拓扑图的方法从 Kinetics-Skeleton 数据集较少的关节点数据中提取出更多的高维特征，使得本文算法相比对比算法在该数据集上的识别准确率提升幅度高于在 NTU-RGB+D 数据集上的识别准确率提升幅度。

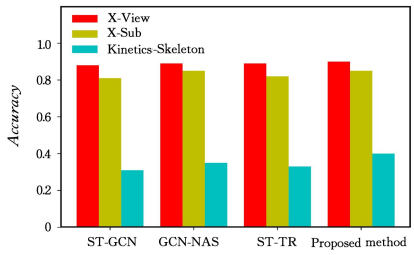

图 4　与不同基线算法的 Top-1 准确率对比图

Fig. 4　Comparison of Top-1 accuracy for the proposed method and different baseline algorithms

4.3.4　消融实验

本节以 Top-1 准确率为评价指标，通过在 Kinetics-Skeleton，NTU-RGB+D(X-Sub)和 NTU-RGB+D(X-View)3 个数据集上的动作识别实验来验证本文提出的动态拓扑图构建模块在人体动作识别过程中所起到的作用和对最终结果的贡献程度，比较结果如图 5 所示。实验结果表明，若移除动态拓扑图构建模块，整个模型缺少了用于计算关节点重要性的 attention 机制以及用于更新节点连接的部分更新机制，使得模型无法选择出某个动作的代表性节点以用于动作识别任务，导致模型性能下降。通过消融实验可证明本文设计的动态拓扑图构建模块提高了动作识别任务的准确率。

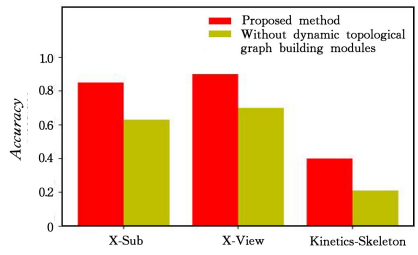

图 5　动态拓扑图构建模块的有效性检验

Fig. 5　Effectiveness of dynamic topological graph building module

结束语　本文提出了基于动态拓扑图的人体骨架动作识别算法，将人体骨架数据编码为关系三元组，基于持续学习的思想，通过解耦合的方式动态构建骨架拓扑图，最后进行人体动作识别，在减小计算代价的同时提升了网络的泛化性。此外，该算法在训练过程中引入了分类准确率评估机制，网络需在每次训练后在之前的所有测试集上取得较好的评估结果，解决了灾难性遗忘问题，提升了基于人体骨架的动作识别算法的性能。然而，本文算法对关节特征变化不明显的动作的识别准确率较低，且采用了按数据类别分批训练的方式，随着动作类别增多，网络学习时间将成倍增长。未来将进一步研究如何降低该算法的时间复杂度以及如何利用深度人体骨架信息增强特征表示能力，以提升对关节特征变化不明显的动作的识别准确率。

参 考 文 献

[1] CHEN Y，TIAN Y，HE M. Monocular Human Pose Estimation：A Survey of Deep Learning-Based Methods[J]. Computer Vision and Image Understanding，2020，192：102897.

[2] SONG S，LAN C，XING J，et al. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data[C]// AAAI Conference on Artificial Intelligence. 2017：4263-4270.

[3] YAN S，XIONG Y，LIN D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// AAAI Conference on Artificial Intelligence. 2018：7444-7452.

[4] XIONG X，MIN W，ZHENG W S，et al. S3D-CNN：Skeleton-Based 3D Consecutive-Low-Pooling Neural Network for Fall Detection[J]. Applied Intelligence，2020，50(10)：3521-3534.

[5] SHI L，ZHANG Y，CHENG J，et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：12026-12035.

[6] ZHANG P，LAN C，ZENG W，et al. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：1112-1121.

[7] DING C Y，LIU K，LI G，et al. Spatio-Temporal Weighted Posture Motion Features for Human Skeleton Action Recognition Research[J]. Chinese Journal of Computers，2020，43(1)：29-40.

[8] TIAN Z Q，DENG C H，ZHANG J W. Human Behavior Recognition Algorithm Based on Skeletal Temporal Divergence Fea-

ture[J]. Journal of Computer Applications, 2021, 41(5): 1450-1457.

[9] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based Action Recognition with Directed Graph Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7912-7921.

[10] TANG Y, TIAN Y, LU J, et al. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5323-5332.

[11] THAKKAR K, NARAYANAN P J. Part-Based Graph Convolutional Network for Action Recognition[J]. arXiv: 1809.04983, 2018.

[12] LI M, CHEN S, CHEN X, et al. Symbiotic Graph Neural Networks for 3D Skeleton-based Human Action Recognition and Motion Prediction[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://ieeexplore.ieee.org/document/9334430.

[13] LI B, LI X, ZHANG Z, et al. Spatio-temporal Graph Routing for Skeleton-based Action Recognition[C]// AAAI Conference on Artificial Intelligence. 2019: 8561-8568.

[14] HADSELL R, RAO D, RUSU A A, et al. Embracing Change: Continual Learning in Deep Neural Networks[J]. Trends in Cognitive Sciences, 2020, 24(12): 1028-1040.

[15] CHEN P H, WEI W, HSIEH C, et al. Overcoming Catastrophic Forgetting by Generative Regularization[J]. arXiv: 1912.01238, 2019.

[16] D'AUTUME C D M, RUDER S, KONG L, et al. Episodic Memory in Lifelong Language Learning[J]. Advances in Neural Information Processing Systems, 2019, 32: 13143-13152.

[17] ROLNICK D, AHUJA A, SCHWARZ J, et al. Experience Replay for Continual Learning[J]. Advances in Neural Information Processing Systems, 2019, 32: 350-360.

[18] LIU L, PU H Y. Real-time LSTM-based Multi-dimensional Features Gesture Recognition[J]. Computer Science, 2021, 48(8): 328-333.

[19] KOU X, LIN Y, LIU S, et al. Disentangle-based Continual Graph Representation Learning[C]// Conference on Empirical Methods in Natural Language Processing. 2020: 2961-2972.

[20] PENG W, HONG X, CHEN H, et al. Learning Graph Convolutional Network for Skeleton-based Human Action Recognition by Neural Searching[C]// AAAI Conference on Artificial Intelligence. 2020: 2669-2676.

[21] PLIZZARI C, CANNICI M, MATTEUCCI M. Skeleton-based Action Recognition Via Spatial and Temporal Transformer Networks[J]. Computer Vision and Image Understanding, 2021, 208: 103219.

XIE Yu, born in 1993, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include graph neural networks and so on.

WANG Wen-jian, born in 1968, Ph.D, professor, Ph.D supervisor, is a distinguished member of China Computer Federation. Her main research interests include machine learning and neural networks.

(责任编辑:喻藜)